Universidade de Aveiro Departamento de Ciências Médicas
2016

Susana Catarina da Costa Ferreira

Estudo de Sinais Epigenéticos no Genoma Humano

Study of the Epigenetic Signals in the Human Genome

Universidade de Aveiro Departamento de Ciências Médicas
2016

Susana Catarina da Costa Ferreira

Estudo de Sinais Epigenéticos no Genoma Humano

Study of the Epigenetic Signals in the Human Genome

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biomedicina Molecular, realizada sob a orientação científica da Professora Doutora Vera Mónica Almeida Afreixo, Professora de Matemática do Departamento de Matemática e membro do Instituto de Biomedicina da Universidade de Aveiro e da Professora Gabriela Maria Ferreira Ribeiro de Moura, Professora Auxiliar Convidada do Departamento de Ciências Médicas e membro do Instituto de Biomedicina da Universidade de Aveiro.

o júri

Presidente

Professora Doutora Odete Abreu Beirão da Cruz e Silva
Professora Auxiliar C/ Agregação, Universidade de Aveiro

Vogais

Professor Doutor João Manuel de Oliveira e Silva Rodrigues
Professor Auxiliar, Universidade de Aveiro

Professora Doutora Vera Mónica Almeida Afreixo
Professora Auxiliar, Universidade de Aveiro (Orientadora)

agradecimentos

À minha orientadora e co-orientadora, a Professora Vera Afreixo e Professora Gabriela Moura, pela sua inegável competência científica na orientação paciente, dedicada e crítica, assim como o seu apoio e disponibilidade ao longo deste percurso.
À minha família pela compreensão e encorajamento, sem a qual esta dissertação não teria sido possível.

palavras-chave

Epigenética, epigenoma, modificação de histonas, metilação do DNA, ncRNAs, epigenoma humano, marcação epigenética, contexto genómico, análise de dados.

resumo

A epigenética pode ser definida pela ocorrência de modificações no genoma, que são herdadas durante a divisão celular, não havendo no entanto modificações directas na sequência do DNA. Estas modificações genómicas são apoiadas em três grandes mecanismos epigenéticos: metilação do DNA, modificação de histonas e pequenos RNAs. Estas diferentes marcas epigenéticas podem ter como função regular a transcrição genética, pois quando existe algum tipo de alteração nestes processos, pode desencadear-se em diversas patologias como o cancro.

Assim, o objectivo principal deste trabalho é estudar os sinais epigenéticos no genoma humano, ou seja, observar se existe dependência entre o contexto e a ocorrência da marcação epigenómica. Para esse efeito foram utilizados os epigenomas das histonas disponíveis na base de dados do NIH Roadmap Epigenomics Mapping Consortium que contêm vários tipos de células e vários tipos de tecidos. No presente estudo são empregues diferentes metodologias estatísticas, nomeadamente testes estatísticos, medidas do tamanho do efeito, análise de resíduos e classificação hierárquica. Com esta análise, comparam-se contextos genómicos da marcação epigenética entre cromossomas e entre epigenomas. Complementando a análise com um cenário de controlo, sem marcação e factorizando pelo teor de CG.

Foi possível identificar uma dependência entre o contexto e a ocorrência de marcação epigenética, sendo possível identificar contextos genómicos específicos para as modificações das histonas.

keywords

Epigenetics, epigenome, histone modification, DNA methylation, ncRNAs, human epigenomes, epigenetic marking, genome context, data analysis.

abstract

Epigenetics can be defined as changes in the genome that are inherited during cell division, however without direct modify the DNA sequence. These genomic changes are supported by three major epigenetic mechanisms: DNA methylation, histone modification and small RNAs. Different epigenetic marks function by regulating gene transcription, because when these processes are altered, this triggers various diseases such as cancer.

Thus, one main objective was to study the epigenetics signals in the human genome, meaning, whether there is dependence observed between the context and the occurrence of epigenomic marking. For this purpose we used histone epigenomes available in the NIH Roadmap Epigenomics Mapping Consortium database that contains various types of cells and various types of tissues. The present study employed different statistical methodologies, namely, statistical tests, effect size measures, residue analysis and hierarchical classification. With this analysis, we compared genomic contexts of epigenetic marking among chromosomes and among epigenomes. Complementing the analysis with a control scenario, without marking and factoring the CG content.

As a result of this study, it was possible to identify one dependency between the context and the occurrence of epigenetic marking and we were able to identify specific genomic contexts in histone modifications.

Index

Index of figures

Index of tables

List of abbreviations

5mC – 5-methylcytosine

A – Adenine

BAM – Binary Sam

BS-seq – Sodium bisulfite

C – Cytosine

Chip-seq – Chromatin Immunoprecitation Sequencing

CpG – Cytosine-phosphate- guanine

DNMTs – DNA methyltransferases

DNA – Deoxyribonucleic acid

DNase – Deoxyribonuclease

E1 – Ubiquitin-activating enzyme

E2 – Ubiquitin-conjugating enzymes

E3 – Ubiquitin ligases

G – Guanine

H1 – Histone protein 1

H2A – Histone protein 2A

H2B – Histone protein 2B

H3 – Histoen protein 3

H4 – Histone protein 4

HDAC – Histone deacetylases

MeDIP-seq – Methylated DNA immunoprecipitation

mRNA – Messenger RNA

miRNA – Micro RNAs

MRE-seq – Methylation-sensitive restriction enzymes

ncRNAs – Non protein coding RNA

NIH – National Institutes of Health

RNA – Ribonucleic acid

RRBS – Reduced representation bisulfite sequencing

PCR – Polymerase chain reaction

piRNA – Piwi interacting RNAs

SRA – Sequence read archives

siRNA – Shirt interfering RNAs

T – Thymine

Treg – Cells T regulatory

U – Uracil

UTR – Untranslated region

WIG – Wiggle formation

1. Introduction

Genetics has been one of the most developed areas of research, particularly after the discovery of the structure of deoxyribonucleic acid (DNA). This area investigates the constitution of genes, processes of inheritance, what physical disorders can be present in the genetic map and how DNA controls the structure, function and behavior of cells and the way it encodes proteins (1).

The genetic language is universal, and is mainly contained in DNA. In turn, DNA is located within almost all human cells, organized into structures that are designated chromosomes. In its composition, DNA has four nitrogenous bases called nucleotides, designated by: adenine (A), guanine (G), thymine (T) and cytosine (C). In addition to these four nucleotides, the DNA is composed by a lateral chain of orthophosphates (phosphoric acid) and deoxyriboses (1).

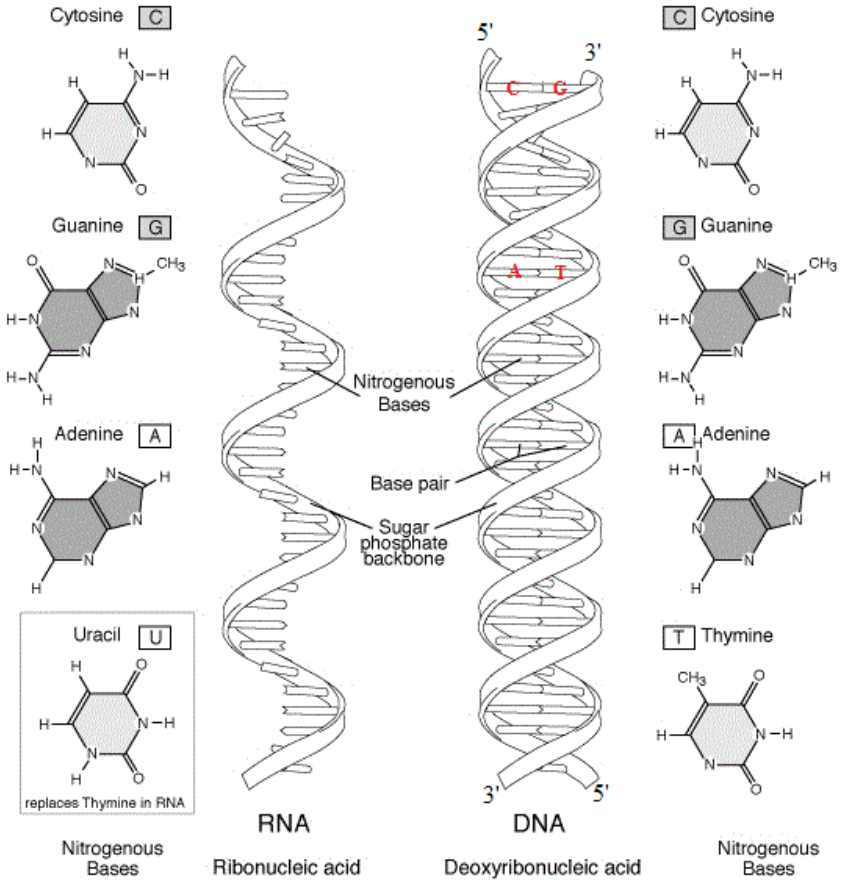


Figure 1: DNA and RNA structure. The DNA molecule has a structure the double helix. The nucleotides are linked through the sugar-phosphate backbone phosphodiester linkages 3'-5. Between the two chains, adenine binds to thymine, and guanine with cytosine constituting the double helix structure. The RNA is a single-stranded molecule containing A, C, G and uracil (U) instead of thymine. Adapted image (2).

As can be seen in figure 1, the DNA molecule has a structure of helix double, forming a spiral. The nucleotides are linked through the sugar-phosphate backbone by phosphodiester linkages 3'-5’ between the two chairs, adenine binds to thymine, guanine to cytosine constituting by wear interactions that allows for the double helix structure. On the other hand, ribonucleic acid (RNA) acts as an intermediate in protein synthesis, acting between DNA and proteins. It is a single-stranded nucleotide polymer, containing A, C, G and uracil (U). Furthermore, it has ribose instead of deoxyribose as sugar constituent (1).

The C+G content it is important and is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine and the C+G pair between to complementary DNA chains is bound by three hydrogen bonds, so DNA with high C+G content is more stable than DNA with low C+G content (1).

This content is associated with the denaturation of DNA, which is the process by which double stranded DNA unwinds and separates into two single strands, through the breaking of hydrogen bonds between complementary bases. With the increase of temperature, the breaking of bonds increases, meaning that the temperature increase acts as a denaturing agent. This melting temperature depends directly on the guanine and cytosine content of the molecule, because the higher the C+G content is the largest numbers of triple bonds have to be broken, requiring a higher temperature for denaturation to occur. After denaturation, if the temperature decrease, there is hybridization of DNA, i.e. re-established went of hydrogen bonding among the complementary bases and return to the double-stranded form (1).

The DNA can be interpreted as the sequence of four nucleotides A, C, T and G, which contain information on hereditary characteristics and are necessary for the continuous production of protein and consequent survival of living beings. We can also have dinucleotides that are composed of two monomers are united nucleotides; the compounds of three nucleotide monomers are known as trinucleotides; and tetranucleotides that are composed of four monomers are united nucleotides.

However, there is a level of gene regulation that does not involve altering DNA sequence. This is called epigenetics, and is one of the most promising and intriguing areas of genetics. Epigenetics is the science that studies the interaction between gene regulation, i.e. how genes are expressed, and its surrounding environment without

involving changes in the DNA sequence level, which may still persist in future generations (3–5).

The cell differentiation is a natural event in every organism, which involves no alteration of DNA sequence. However, each cell type has its own gene expression pattern. All cells in an organism share the same genome (except B lymphocytes), each cell type possesses different kinds of epigenetic signature, and each has a cell-type specific epigenome.

“Epigenetic memory” is a new but distinct term, that refers to the transmission of a gene expression state through multiple cell generations in the absence of initiation signs and genetic variation and could be propagated by various epigenetic mechanisms, such as, DNA methylation, histone modifications and replacement of histone variants (3–5). The inheritance of these epigenetic marks, from mother to daughter cells, is crucial for the maintenance of a cell differentiation state.

In other words, epigenetic has to do with changing the whole genome regulatory activity without involving changes in the DNA sequence. This regulatory activity can be resumed in the epigenome, which is a kind of map that overlays the map of the genome, with epigenetic makes that turn on or off genes, increasing or reducing its activity.

This can only be studied through genomics, which refers to genetic sequencing and analysis of the global genome of an organism. Wherein, the genome is the integer index of DNA that is present within a cell of an organism (6,7). Genomics aims to determine complete sequences of DNA and draw their genetic trait, in order to help understand its functions and dysfunctions, and to study all genes, involving DNA, messenger RNA (mRNA), and proteome at both the cell level or tissue level (6,7).

An important note is that epigenome is not static as the genome, meaning, that the epigenome can be dynamic, influenced by environmental factors and extracellular stimuli, and change rapidly in response to these factors (6,7).

As we will see, epigenetics can be regulated by three mechanisms: DNA methylation, histone modification and small RNAs. However, in this work, we have studied epigenetic signals present in the human genome, focusing mainly on the epigenetic regulation related to histone modification.

1.1. Histone modifications and chromatin structure

At the level of chromatin, it exists a mechanism of epigenomic control, since nuclear DNA is associated with histones and other proteins in a complex structure referred to as chromatin. As can be seen in figure 2, DNA is wrapped around two copies of each of the four core histone proteins H3, H4, H2B, and H2A (see example in figure 3), to form the nucleosome which is the fundamental repeating unit of chromatin. The nucleosomes are condensed into higher order structures, which in turn form the chromatin (8–10). This will be necessary for efficient packaging of the DNA into the nucleus of the cell. However, because when DNA is compacted into this structure, its accessibility becomes greatly limited, it serves as a mechanism by which the cell protects DNA from external damage but it also regulates DNA mediated processes, such as transcription, DNA replication, DNA repair and chromosome segregation (8–10).

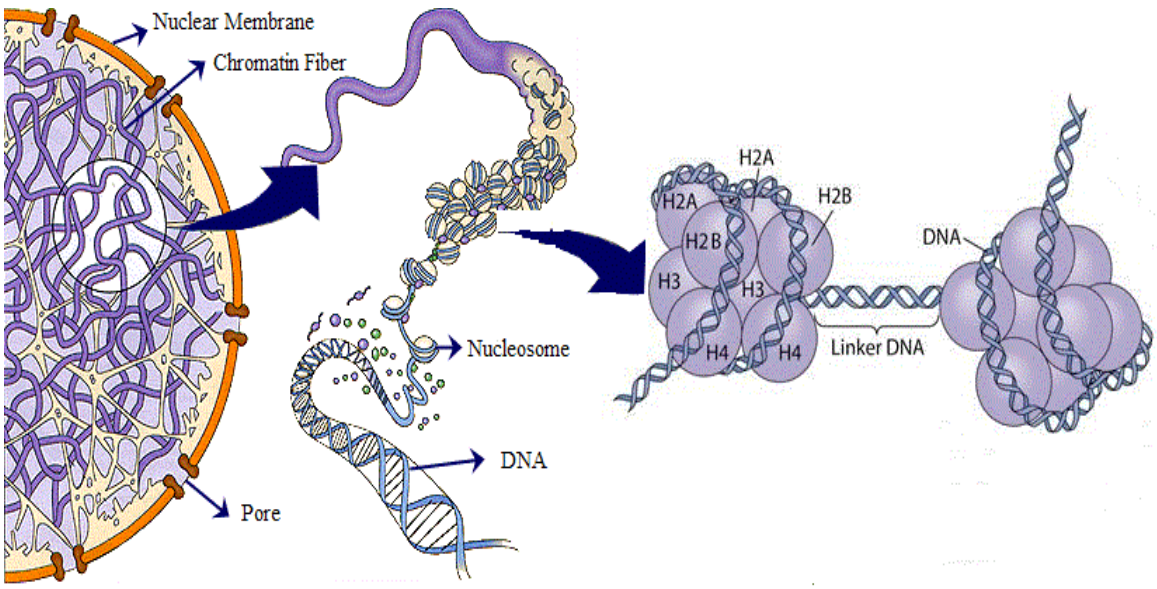


Figure 2: Structure of chromatin. DNA is wrapped around two copies of each of the four core histone proteins H3, H4, H2B, and H2A, to form the nucleosome which is the fundamental repeating unit of chromatin. The nucleosomes are condensed into higher order structures, which in turn form the chromatin. Adapted image (11,12).

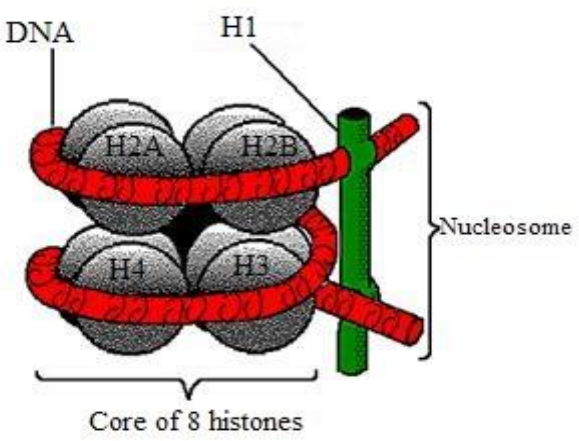


Figure 3: Types of histones: H1, H2A, H2B, H3 and H4. The four core histone proteins H3, H4, H2B, and H2A, form the nucleosome. Histone H1 binds to DNA at the end of each nucleosome and is important for spacing the internucleosomal region. H1 is also known as H5, the basic structure of chromatin consists of 200 pairs of bases of DNA attached to an octamer of histones (H2A, H2B, H3 and H4); H2A and H2B have much lower molecular weight than histone H1 and are considered rich in lysine; H3 and H4 are rich in arginine. Two histones from each class (H2A, H2B, H3 and H4) aggregate to form a nucleosome, together with DNA. Histone H1 is required for DNA-histone complexes to form a fibre, thereby winding the DNA even more effectively. Adapted image (13).

So, these histone proteins can influence chromatin organization and regulate many DNA-templated processes, through their chemical modification patterns (acetylation, methylation, sumoylation, and ubiquitylation) (8,14). Changes on the histone modification status may be associated with active or inactive chromatin. In addition, the combinatorial nature of various histone modifications occurring at different times during development, and at specific sites within histones, provides additional levels of regulation and complexity to the epigenome (9).

Regarding the structure of each histone, each one is composed of a conserved globular histone-fold domain, with extended N and C-terminal tails. The globular domains of the histone proteins form the nucleosome core, around which 146 base pairs of DNA are wrapped (10). The histone tails serve as the modification site as explain abowe (10).

More recently, the post translational modifications done to histones include lysine and arginine methylation, lysine acetylation, serine and threonine phosphorylation, monoubiquitylation, sumoylation and proline isomerization, as described in detail in annex 1 (8,10).

How many of these modifications exert some biological effect, and how the addition or removal of many of these modifications is regulated is still unclear. These modifications may directly affect chromatin structure, such as altering histone-DNA or

internucleosomal interactions, whereas others may act indirectly, recruiting or preventing the binding of specific proteins to chromatin (8). Also, in some cases, recognition of specific histone modifications by various effector proteins is thought to mediate specific biological processes such as gene activation or gene repression (8,15). Furthermore, alteration of histone charges by modifications, such as acetylation and phosphorylation are thought to induce localized structural changes in chromatin, allowing protein factors to access the DNA. Thus, it has been observed that misregulation of the addition and/or removal of these modifications, or of the enzymes that catalyse them, are implicated in human disease, including various cancers (9).

In addition, it is important to understand the notation of histone modifications, since a wide array of them has been identified and a standardized nomenclature has been proposed. The Brno nomenclature (example in figure 4) was created by a consortium of European laboratories to standardize notation of histones and histone modifications (15).

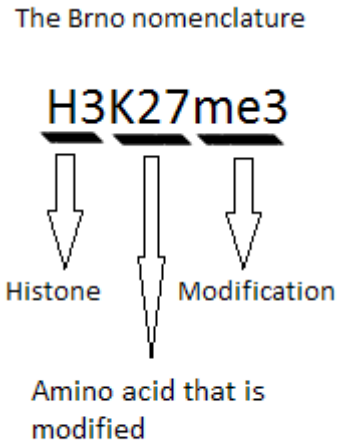


Figure 4: The Brno nomenclature for histone modifications. The histone protein (H3, H4, H2A, or H2B) is indicated first, followed by the amino acid that is modified (ex. "K27" represents lysine 27). This is followed by the type of modification that is observed (ex. "me3" represents tri-methylation).

The modification of histone proteins is another possible mechanism of epigenetic memory. Histone proteins are subjected to various types of post translational modifications, as previously mentioned, and these modifications are associated with different cellular functions, such as transcriptional/translational regulation, DNA replication/repair and nucleosome compaction. However, the mechanism of replication of histone modification after cell division is less understood, and it is not yet clear how

the histone modifications, and hence such an epigenetic memory contained in a mother cell, can be inherited precisely by its daughter cells (9).

However, it is important to understand that these epigenetic marks can persist and influence a gene state, making it active or silent, even after many cell divisions.

1.2. DNA methylation

The methylation of cytosine residues in DNA molecules was the first epigenetic discovery. DNA methylation is linked to transcriptional silencing, and is important for gene regulation, development, and tumorigenesis (8,16,17).

This epigenetic regulation mechanism relies on the activity of DNA methyltransferases (DNMTs), that are responsible for the transfer of a methyl group to DNA and are essential for mammalian development (18). Nearly all DNA methylation occurs on cytosine residues that are located side by side with guanine residues, forming cytosine-phosphate-guanine (CpG) dinucleotides, which usually appear heavily repeated in genomic sequences and are called CpG islands. Usually these regions are preferably found in non-coding regions of the genome. In coding ones, they are almost exclusively in 5' ends, like promoters, 5' untranslated regions (5' UTR) or the exon 1 of human genes, and their methylation degree might prevent gene expression (8). However, there are other genomic regions affected by DNA methylation, located nearby CpG islands, but with less CpG dinucleotides, termed CpG shores. Indeed, these regions may also regulate gene expression.

Methylation is a reversible reaction, and so demethylation might also take place, promoting the activation of gene transcription. This process is mediated by the action of TET family proteins or by GADD45 family members (14).

DNA methylation may prevent expression, either directly through transcriptional activators' obstruction, or indirectly by recruitment of methylcytosine-binding proteins. Furthermore, these may promote the enrolment of more DNMTs and histone deacetylases (HDAC), which might result in additional chromatin alterations, that will further repress transcription (14).

In mammalian cells, methylation of cytosine residues is catalysed by DNMTs, including DNMT1, DNMT3a and DNMT3b (8,19). Wherein, the DNMT3a and DNMT3b are involved in de novo methylation, which may occur in embryonic stem cells or cancer cells (19). While the DNMT1 maintains the genomic methylation state

by specifically recognizing and methylating hemimethylated CpG dinucleotides during DNA replication (20–22).

The association between DNA methylation and disease is that aberrant patterns of DNA methylation influence many aspects of disease processes, through altered gene expression profiles, particularly in many human tumours (16,17,23). Cancers have the unique property of being globally hypomethylated, which alters the chromatin architecture, leading to the inappropriate activation of oncogenes. In contrast, specific hypermethylation and hence silencing of tumour suppressor genes is recognized as a hallmark of many types of cancer cells (8,24,25).

DNA methylation can also be a plausible explanation for the phenomenon of epigenetic memory, because it is associated with gene silencing. During DNA replication, each daughter cell will have one parental DNA strand, which carries the mother cell's DNA methylation pattern; and one newly synthesized DNA strand, which is at first unmethylated (3). As previously mentioned, these hemi-methylated sites are the preferential substrates of DNA methyltransferase (DNMT), DNMT1 and consequently the DNA methylation pattern in daughter cells will be faithfully replicated from the mother cell, hence explaining the epigenetic memory (8).

In the "Roadmap Epigenomics Consortium" that is the database we use, DNA methylation is assayed by sequencing DNA using three different methods: BS-seq; MeDIP-seq; and MRE-seq.

- BS-seq

A gold-standard technology for detection of DNA methylation is bisulfite genomic sequencing, because it provides a qualitative, quantitative and efficient approach to identify 5-methylcytosine (5mC) at single base-pair resolution. Frommer et al (1992), was the first to introduce this method which is based on the finding that the amination reactions of cytosine and 5-methylcytosine proceed with very different consequences after the treatment with sodium bisulfite (26,27).

The process consists of cytosines in single-stranded DNA that will be converted into uracil and recognized as thymine in the subsequent polymerase chain reaction (PCR) for amplification and sequencing. However, 5mCs are immune to this conversion and remain as cytosines allowing 5mCs to be distinguished from unmethylated cytosines after sequencing. To determine the methylation status in the loci of interest by using specific methylation primers after the bisulfite treatment is necessary the PCR

process. The actual methylation status can be determined either through direct PCR product sequencing (detection of average methylation status) or sub-cloning sequencing (detection of single molecules distribution of methylation patterns), as can be seen in figure 5. Furthermore bisulfite sequencing analysis (see example in figure 6) can identify the DNA methylation status along the DNA single strand, and can also detect the DNA methylation pattern of DNA double strands since the converted DNA strands are no longer self-complementary and the amplification products can be measured individually (27,28).

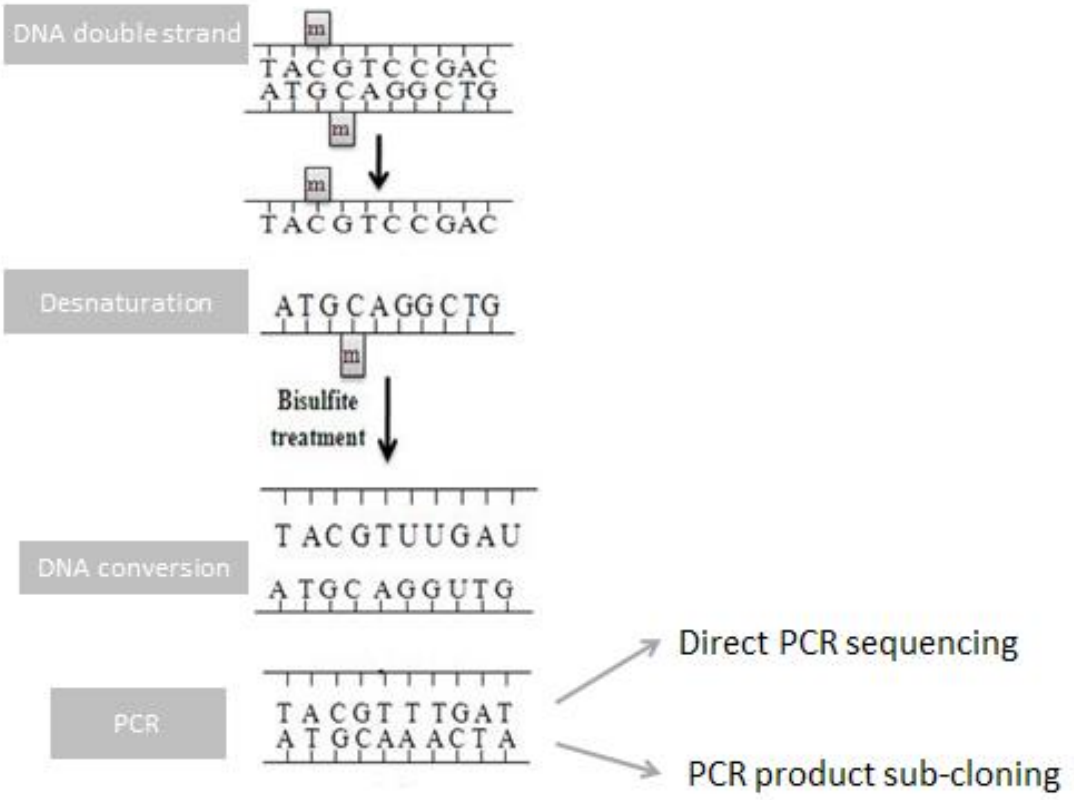


Figure 5: Principles of methylation analysis using bisulfite genomic sequencing. After treatment with sodium bisulfite, unmethylated cytosine residues are converted to uracil whereas 5mC remains unaffected. After PCR amplification, uracil is converted to thymine. The DNA methylation status can then be determined by direct PCR sequencing or cloning sequencing. Adapted image (27).

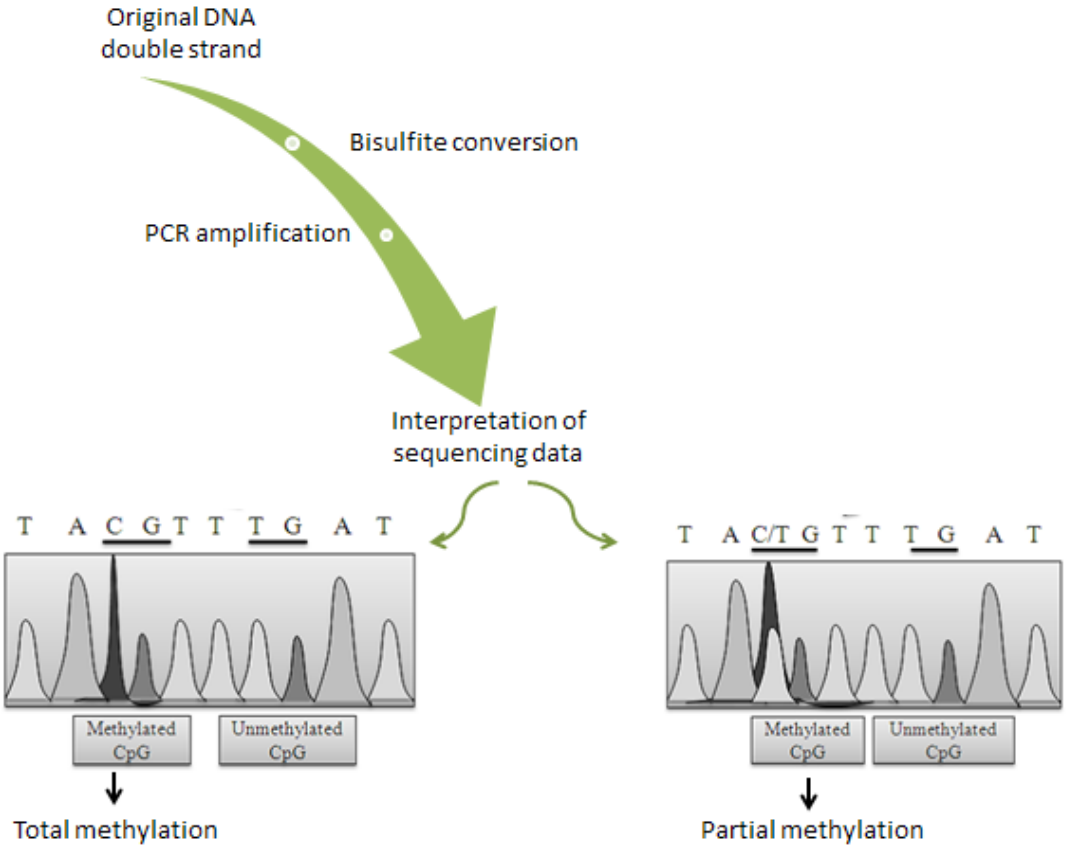


Figure 6: Interpretation of methylation sequencing results. DNA methylation status can be interpreted by comparing the sequencing results and the original DNA sequence. Fundamentally, all unmethylated C convert to T and the presence of a C-peak indicates the presence of 5mC in the genome. But, if both C-and T-peaks appear, this indicates that partial methylation or potentially incomplete bisulfite conversion has occurred. The proportion of 5mC to C can be interpreted by analysing the relative square area of these two bands. Adapted image (27).

Bisulfite-based DNA methylation analyse are more quantitative, sensible, efficient, and allows for a wide spectrum of sample analyses, compared with other DNA methylation approaches, some of them based on the sensitivity of restriction enzymes that can specifically recognize methylated cytosines within their cleavage recognition sites (28–33). Therefore, this method is the golden standard for DNA methylation analysis and has been widely used in various research and clinical settings.

- MeDIP-seq and MRE-seq

These two methods (as we can see in figure 7) have their own advantages and disadvantages, and can be applied independently or jointly. However, Li D. et al (2011), suggest that by integrating these two complementary technologies, we can effectively improve the coverage and resolution of the DNA methylomes produced, and improve the accuracy of detection of differentially methylated regions (34).

The MRE-seq (see in example in figure 8) gives DNA methylation estimates at single CpG resolution, but its coverage is low due to the limit of CpG containing recognition sites. As to MeDIP, it has an important advantage over enzymatic digestion-based methods which is the lack of bias for a specific nucleotide sequence, other than CpGs. However, the relationship of enrichment to absolute methylation levels is confounded by variables such as CpG density. Another limitation of MeDIP-seq is its lower resolution (~150bp) compared to MRE-seq or bisulfite-based methods since one or more of the CpGs in the immunoprecipitated DNA fragment can be the responsible for the antibody binding (28,34,35).

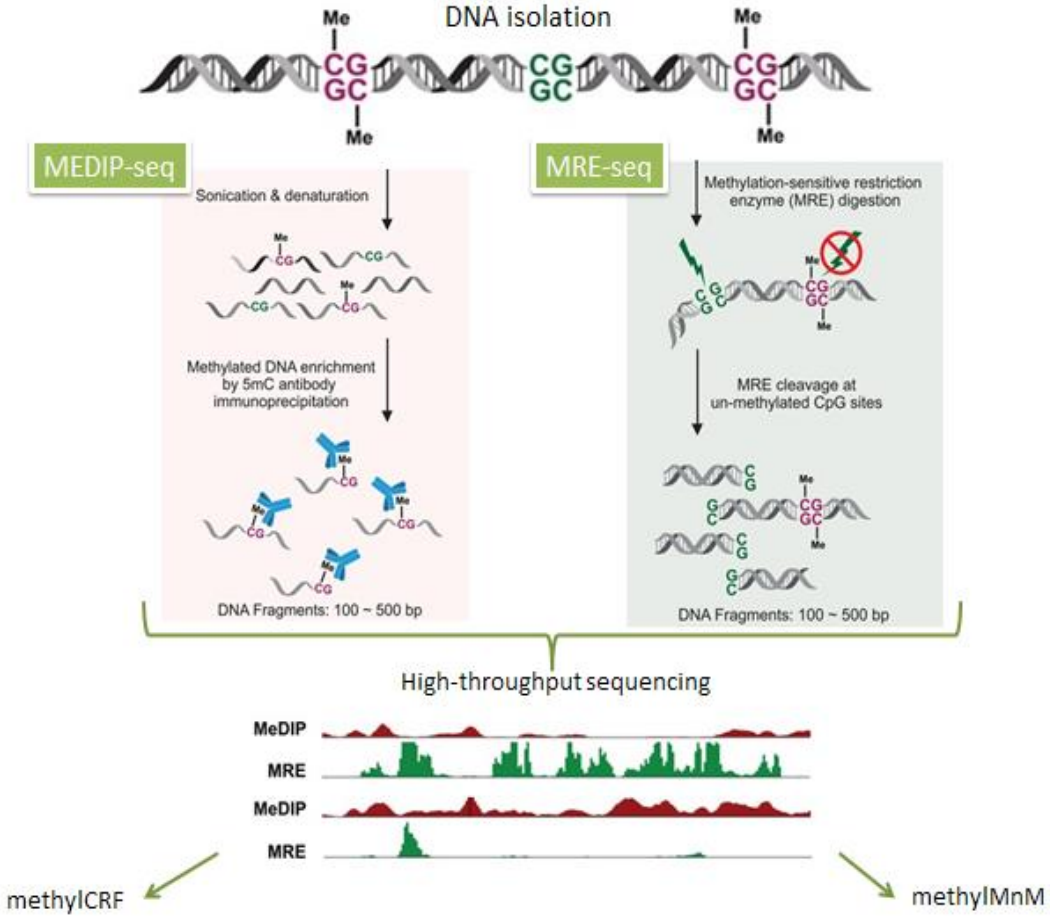


Figure 7: Workflow of combined MeDIP-seq and MRE-seq. Genomic DNA is frist isolated and purified. For MeDIP-seq genomic DNA is sonicated to a specific size range, and a monoclonal anti-5'methylcytosine antibody is used to enrich the sample in methylated DNA fragments. This immunoprecipitated DNA fragments are sequenced and mapped back to the reference genome assembly to discover methylated regions. For MRE-seq genomic DNA is sonicated to a specific size range, and a monoclonal anti-5'methylcytosine antibody is used to enrich the sample in methylated DNA fragments. This immunoprecipitated DNA fragments are sequenced and mapped back to the reference genome assembly to discover methylated regions. Both the MeDIP-seq as MRE-seq data can then be integrated by applying methylCRF, which transform enrichment-based DNA methylation data to methylation level at single CpG resolution across the genome. To compare two samples and detect differentially methylated regions, M&M is applied in a region-specific fashion, for example. Adapted image (34).

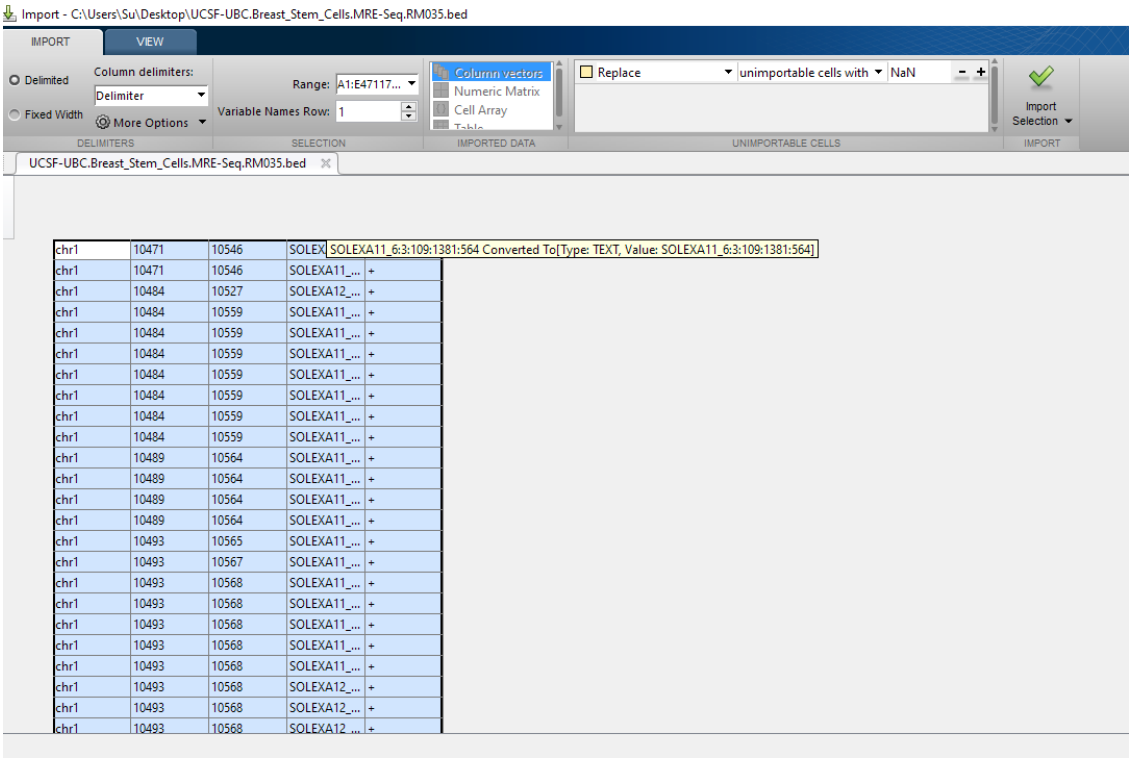


Figure 8: Example of a MRE-seq file, which will be used in this work, related to breast stem cells. First column shows the chromosome (chromosome 1), second column the first position and third column the last position of DNA restriction site. Image taken from MATLAB.

1.3. Small RNAs

There is growing evidence that dynamic changes on chromatin, chromosomes and nuclear architecture are regulated by RNA signalling. Indeed, through small non protein coding RNAs (ncRNAs), regulation can occur at some of the most important levels of genome function, including chromatin structure, chromosome segregation, transcription, RNA processing, RNA stability, and translation (36,37).

While the precise molecular mechanisms are not yet well understood, they appear to involve the differential recruitment of a hierarchy of generic chromatin modifying complexes, and DNA methyltransferases to specific loci by RNAs during differentiation and development. A significant fraction of the genome-wide transcription of ncRNAs may be involved in this process, comprising a previously hidden layer of intermediary genetic information that underpins developmental ontogeny and the differences between species, ecotypes and individuals (36,37).

In addition, RNA-directed regulatory processes may also transfer epigenetic information not only within cells but also among cells and organ systems, as well as across generations.

Therefore, small ncRNAs have an important role in regulating gene expression and many classes of small RNAs have been described. There is general recognition of three main categories the small ncRNAs, differing on various aspects of their origins, structures, associated effector proteins, and biological roles: short interfering RNAs (siRNA), microRNAs (miRNA) and piwi interacting RNAs (piRNA), which are described in detail in annex 2.

The processes directed by these small RNAs will confer resistance to a variety of cellular insults, such as viral infection, and prevent random transposition events within the genome. In addition, small RNAs have been shown to be important for directing other epigenetic processes, such as DNA methylation and chromatin modification, as mentioned above (28,36,37).

These three mechanisms of epigenetic regulation (DNA methylation, histone modifications and chromatin structure, and finally small RNAs), contribute to the epigenome, which is like a computer program that regulates the functioning of the genes themselves. The distribution of methylated DNA, histone modifications, and ncRNA expression may not only be specific to a particular organism, but it will be specific to a particular tissue, or even one particular cell type. Misregulation of these epigenetic events has been observed in various cancers and human diseases. So, understanding how the epigenome contributes to gene regulation will give us greater insight into the thin frontier between human health and disease.

1.4. General objectives

The work described in this dissertation aimed to study the epigenetic signals of the human genome, to observe the dependence between the context and the occurrence of epigenetic marking and, if possibly, to identify specific contexts related to genomic modification of histones.

Thus, the main objectives of this work were:

- Obtain the frequency of oligonucleotides, in marked regions and unmarked regions.
- Using statistical methods, such as statistical tests, residue analysis, effect size measures and hierarchical classification, comparing the genomic

context of epigenetic marking among chromosomes and among epigenomes.

- Create a control dataset to complement the analysis.
- Statistical evaluation.

1.5. Dissertation structure

As mentioned earlier, this work consists on the determination of genomic contexts specific of epigenetic markings of histone modifications. To do this it was necessary to developed some R scripts in program R (described in Chapter 3), so as to obtain the frequencies of oligonucleotides.

Firstly, it was necessary to obtain of histone modifications in public databases (described in Section 2). Then, we obtained the frequency of oligonucleotides of those fragments (described in Chapter 3) and then we summed the total of such frequencies. In order to obtain these results, we used the statistical methods described in Chapter 3.

This dissertation is divided into three chapters and annexes:

- In chapter 1 it is introduced the theme of the dissertation and the general objectives are presented.
- In chapter 2 ("Roadmap Epigenomics Consortium") is presented the database we use, as well as a description of the available and used files.
- In chapter 3 ("Analysis of epigenomes") it is presented a description of the tools used as well as the statistical procedure. This chapter will also present the results obtained.
- In chapter 4 ("Discussion") the main conclusions obtained is this work are presented and also ideas for future work.
- Annex 1 ("Modifications of histones") presents a detailed description of each histone modification.
- Annex 2 ("The three main categories of small ncRNAs") presents the description of each category the small ncRNAs.
- Annex 3 ("The various tissues present in Consortium") presents the description of each cell line used in the study.
- Annex 4 ("Characteristics of each histone modification") presents the main characteristics of each histone modification studied in this dissertation

- Annex 5 (“R scripts”) describes each R script used and presents their functions.
- Annex 6 ("Histone union and histone intersection") presents the results obtained for histone union and histone intersection, using the chi-square test and measure of association Cramer's V.
- Annex 7 ("Comparing to the histone with each control") presents the results of comparing each histone modification with the control through the chi-square test and the measure of association Cramer's V.
- Annex 8 ("Heatmaps") presents examples of heatmaps about the genomic context of epigenetic marking.

2. Roadmap Epigenomics Consortium

2.1. Project description

The NIH Roadmap Epigenomics Mapping Consortium was designed, in 2008, to produce a database of human epigenomic data in order to encourage biology and research associated with the disease (38,39).

These epigenomic maps, "offer broad insight into genome regulation and are generally available for diverse cell populations". The Consortium includes genomic maps for: (39):

i. DNA methylation: assayed by sequencing DNA that has been treated with sodium bisulfite (BS-seq), or enriched in methylated cytosines by methylcytosine pull down (methylated DNA immunoprecipitation (MeDIP-seq)) or methylation-sensitive restriction reactions (MRE-seq).

ii. Histone modifications: assayed by sequencing DNA enriched by chromatin immunoprecipitation with modification-specific histone antibodies (ChIP-seq).

iii. Chromatin accessibility: assayed by sequencing DNase I cleavage sites in nuclear chromatin.

iv. RNA expression: assayed by sequencing mRNAs or size-selected small RNA fractions to high depths. These expression data are intended to augment and illuminate the functional output of the epigenomic profiles.

The Consortium investigate a diverse set of cell and tissue models (Table 1), These cells and tissues were prioritized on the basis of broad scientific and biomedical interest. All cell lines are described in detail in annex 3.

Adrenal – Fetal ▪ Fetal, Adrenal Gland
Brain ▪ Brain Anterior Caudate ▪ Brain Cingulate Gyrus ▪ Brain Hippocampus Middle ▪ Brain Angular Gyrus ▪ Brain Inferior Temporal Lobe ▪ Brain Germinal Matrix ▪ Brain Mid Frontal Lobe ▪ Brain Substantia Nigra ▪ Neurosphere Cultured Cells Cortex Derived

- Neurosphere Cultured Cells Ganglionic Eminence Derived
- Brain Cerebellum

Brain – Fetal

- Fetal, Brain
- Fetal Spinal Cord

Breast

- Breast Stem Cells
- Breast Myoepithelial Cells
- Breast Luminal Epithelial Cells
- Breast vHMEC

ES Cells

- H1
- H9
- HUES1
- HUES3
- HUES6
- HUES8
- HUES9
- HUES13
- HUES28
- HUES44
- HUES45
- HUES48
- HUES49
- HUES53
- HUES62
- HUES63
- HUES64
- HUES65
- HUES66
- ES-I3 Cell Line
- ES-WA7 Cell Line
- UCSF-4Star

ES – Derived Cells

- H1 Derived Embryoid Body Cultured Cells
- H9 Derived Embryoid Body Cultured Cells
- HUES1 Derived Embryoid Body Cultured Cells
- HUES3 Derived Embryoid Body Cultured Cells
- HUES6 Derived Embryoid Body Cultured Cells
- HUES45 Derived Embryoid Body Cultured Cells
- H1 Derived Neuronal Progenitor Cultured Cells
- H9 Derived Neuronal Progenitor Cultured Cells
- H9 Derived Neuron Cultured Cells
- H1-BMP4
- hESH1 derived mesenchymal
- hESH1 derived mesendoderm
- hESC Derived CD184+ Endoderm Cultured Cells
- hESC Derived CD56+ Mesoderm
- hESC Derived CD56+ Ectoderm Cultured Cells

Exocrine – Endocrine

- Adrenal Gland

▪ Pancreas ▪ Spleen ▪ Thymus
Fat – Adult ▪ Mesenchymal Stem Cell Derived Adipocyte Cultured Cells ▪ Adipose Nuclei ▪ Adipose Derived Mesenchymal Stem Cell Cultured Cells ▪ Adipose Tissue
GI – Adult ▪ Liver, Adult ▪ Stomach Smooth Muscle ▪ Duodenum Mucosa ▪ Duodenum Smooth Muscle ▪ Pancreatic Islets ▪ Colonic Mucosa ▪ Rectal Mucosa ▪ Rectal Smooth Muscle ▪ Colon Smooth Muscle ▪ Esophagus ▪ Gastric ▪ Sigmoid Colon ▪ Small Intestine
GI – Fetal ▪ Fetal, Intestine Large ▪ Fetal, Intestine Small ▪ Fetal, Stomach
GU – Adult ▪ Kidney, Adult ▪ Bladder
Heart – Adult ▪ Aorta ▪ Heart ▪ Left Ventricle ▪ Right Atrium ▪ Right Ventricle
Heart – Fetal ▪ Fetal, Heart
Hematopoietic Stem ▪ CD34, Primary Cells ▪ CD34, Mobilized Primary Cells ▪ CD34, Cultured Cells

IPs Cells

- IPS 4.7
- IPS 6.9
- IPS-11a
- IPS-11b
- IPS-11c
- IPS-15b
- IPS-17a
- IPS-17b
- IPS-18a
- IPS-18b
- IPS-18c
- IPS-19.7
- IPS-19.11
- IPS-20b
- IPS-27b
- IPS-27e

Kidney – Fetal

- Fetal, Kidney
- Fetal, Kidney Left
- Fetal, Kidney Right
- Fetal, Renal Cortex
- Fetal, Renal Cortex Left
- Fetal, Renal Cortex Right
- Fetal, Renal Pelvis
- Fetal, Renal Pelvis Left
- Fetal, Renal Pelvis Right

Lung – Adult

- Lung

Lung – Fetal

- Fetal, Lung
- Fetal, Lung Left
- Fetal, Lung Right
- Fetal, Lung Fibroblast (IMR90)

Muscle - Adult

- Skeletal Muscle
- Muscle Satellite Cultured Cells
- Psoas Muscle

Muscle - Fetal

- Fetal, Muscle Arm
- Fetal, Muscle Back
- Fetal, Muscle Leg
- Fetal, Muscle Lower Limb
- Fetal, Muscle Upper Limb
- Fetal, Muscle Trunk
- Fetal, Muscle Upper Trunk

Placenta – Fetal

- Fetal, Placenta
- Placenta Amnion
- Placenta Basal Plate
- Placenta Chorion Smooth

- Placenta Trophoblast Primary Cells
- Placenta Villi

Reproductive – Adult

- Ovary
- Testis Spermatozoa Primary Cells

Reproductive – Fetal

- Fetal Ovary

Skin – Fetal

- Fibroblasts Fetal Skin Abdomen
- Fibroblasts Fetal Skin Biceps Left
- Fibroblasts Fetal Skin Biceps Right
- Fibroblasts Fetal Skin Quadriceps Left
- Fibroblasts Fetal Skin Quadriceps Right
- Fibroblasts Fetal Skin Upper Back
- Fibroblasts Fetal Skin Upper Back

Spleen – Fetal

- Fetal, Spleen
- Fibroblasts Fetal Skin Back

Stromal – Connective

- Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells
- Chondrocytes from Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells
- Primary Fibroblast
- Fetal Skin
- Foreskin Fibroblast Primary Cells
- Foreskin Keratinocyte Primary Cells
- Foreskin Melanocyte Primary Cells
- Breast Fibroblast Primary Cells

Thymus – Fetal

- Fetal, Thymus

White Blood

- CD3+ Total Unmobilized
- CD3+ Total Mobilized
- CD8, Primary Cells
- CD8, Naive Primary Cells
- Mobilized CD8 Primary Cells
- CD4, Primary Cells
- CD4, Mobilized Primary Cells
- CD4, Memory Primary Cells
- CD4+ CD25- CD45RA+ Naive Primary Cells
- CD4, Naive Primary Cells
- CD4+ CD25+ CD127- Treg Primary Cells
- Treg Primary Cells
- Th17 Primary Cells
- CD4+ CD25- IL17+ PMA-Ionomcyin stimulated Th17 Primary
- CD4+ CD25- CD45RO+ Memory Primary Cells
- CD4+ CD25- IL17- PMA-Ionomycin stimulated MACS purified Th Primary Cell
- CD4+ CD25- Th Primary Cells
- CD4+ CD25int CD127+ Tmem Primary Cells
- CD14, Primary Cells
- CD56, Primary Cells
- CD8, Mobilized Cells

▪ CD15, Primary Cells ▪ CD19, Primary Cells ▪ CD20, Primary Cells ▪ CD56, Mobilized Cells ▪ Peripheral Blood Mononuclear Primary Cells ▪ CD4+ CD25- CD45RA+ Naive Primary Cells ▪ CD4+ CD25- CD45RO+ Memory Primary Cells

Table 1: Diverse set of cell and tissue models available in the Roadmap Database.

2.2. Description of data

Regarding the files found in the NIH Roadmap Epigenomics Mapping Consortium database, it is important to know that DNA sequencers are devices that read a DNA sample and generate an electronic file with symbols representing the sequence of nitrogenous bases – A, C, G, and T - in the sample. Each sequencing technology has a different strategy, but normally we can identify commons steps between all sequencers: sample preparation, amplification and preparation of the sequencing library (40), sequencing reaction and data analysis.

In the consortium database we can find four types of files:

- Sequence Read Archives (SRA): raw DNA fragments.
- Wiggle format (WIG): fragments already mapped on the reference genome.
- Binary SAM (BAM): fragments already mapped on the reference genome.
- BED: annotation file.

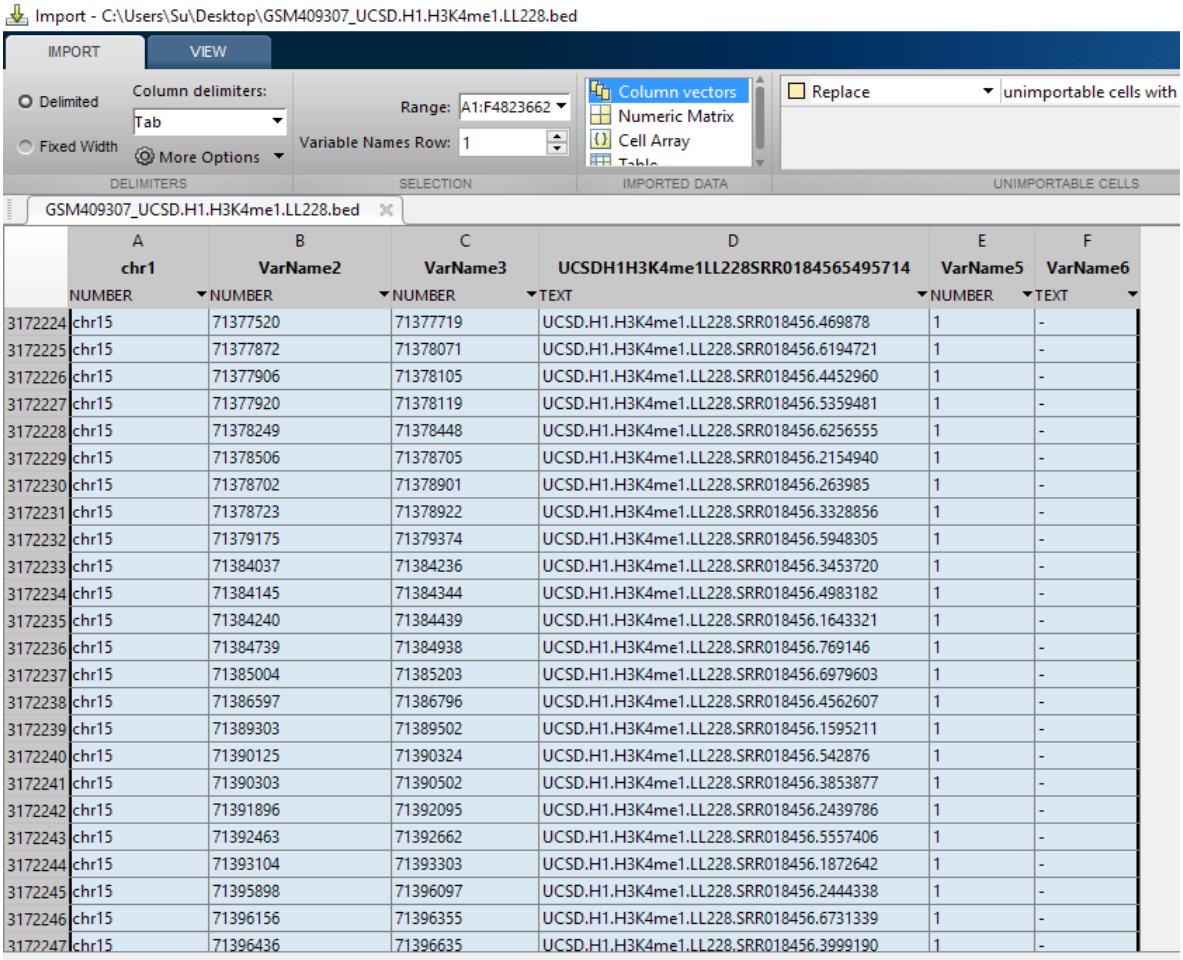
The linear sequence of bases which make up the genome is only the first level of genetic information, where genes and gene expression regulatory elements are encoded. To identify the position of these elements, we use annotation files, which often are nothing more than text files that associate genomic coordinates to genetic characteristics or features.

The BED format (see example in figure 9) is one of these annotation file formats. It was created by UCSC to describe and annotate a genome and is the type of file that we will use in this work. It is a format that can be read by the UCSC browser and has various fields that describe the graphical representation of the information. It is a text format, with columns separated by tab characters representing breaks in the genome associated with the annotation or feature. There are three variations of bed format, the standard version, the version bedDetail/bedGraph, and a binary version called bigBed (40).

In common, all bed formats have the following columns:

1) chrom: chromosome scaffold or name.
2) chromStart: beginning of an annotation position (starting at 0)
3) chromEnd: last position of an annotation.

In describing the features that it contains, a bed file may have a header that begins with the word track and depicts other attributes used for the graphic representation of regions, in the genome browser.



	A	B	C	D	E	F
	chr1	VarName2	VarName3	UCSDH1H3K4me1LL228SRR0184565495714	VarName5	VarName6
	NUMBER	NUMBER	NUMBER	TEXT	NUMBER	TEXT
3172224	chr15	71377520	71377719	UCSD.H1.H3K4me1.LL228.SRR018456.469878	1	-
3172225	chr15	71377872	71378071	UCSD.H1.H3K4me1.LL228.SRR018456.6194721	1	-
3172226	chr15	71377906	71378105	UCSD.H1.H3K4me1.LL228.SRR018456.4452960	1	-
3172227	chr15	71377920	71378119	UCSD.H1.H3K4me1.LL228.SRR018456.5359481	1	-
3172228	chr15	71378249	71378448	UCSD.H1.H3K4me1.LL228.SRR018456.6256555	1	-
3172229	chr15	71378506	71378705	UCSD.H1.H3K4me1.LL228.SRR018456.2154940	1	-
3172230	chr15	71378702	71378901	UCSD.H1.H3K4me1.LL228.SRR018456.263985	1	-
3172231	chr15	71378723	71378922	UCSD.H1.H3K4me1.LL228.SRR018456.3328856	1	-
3172232	chr15	71379175	71379374	UCSD.H1.H3K4me1.LL228.SRR018456.5948305	1	-
3172233	chr15	71384037	71384236	UCSD.H1.H3K4me1.LL228.SRR018456.3453720	1	-
3172234	chr15	71384145	71384344	UCSD.H1.H3K4me1.LL228.SRR018456.4983182	1	-
3172235	chr15	71384240	71384439	UCSD.H1.H3K4me1.LL228.SRR018456.1643321	1	-
3172236	chr15	71384739	71384938	UCSD.H1.H3K4me1.LL228.SRR018456.769146	1	-
3172237	chr15	71385004	71385203	UCSD.H1.H3K4me1.LL228.SRR018456.6979603	1	-
3172238	chr15	71386597	71386796	UCSD.H1.H3K4me1.LL228.SRR018456.4562607	1	-
3172239	chr15	71389303	71389502	UCSD.H1.H3K4me1.LL228.SRR018456.1595211	1	-
3172240	chr15	71390125	71390324	UCSD.H1.H3K4me1.LL228.SRR018456.542876	1	-
3172241	chr15	71390303	71390502	UCSD.H1.H3K4me1.LL228.SRR018456.3853877	1	-
3172242	chr15	71391896	71392095	UCSD.H1.H3K4me1.LL228.SRR018456.2439786	1	-
3172243	chr15	71392463	71392662	UCSD.H1.H3K4me1.LL228.SRR018456.5557406	1	-
3172244	chr15	71393104	71393303	UCSD.H1.H3K4me1.LL228.SRR018456.1872642	1	-
3172245	chr15	71395898	71396097	UCSD.H1.H3K4me1.LL228.SRR018456.2444338	1	-
3172246	chr15	71396156	71396355	UCSD.H1.H3K4me1.LL228.SRR018456.6731339	1	-
3172247	chr15	71396436	71396635	UCSD.H1.H3K4me1.LL228.SRR018456.3999190	1	-

Figure 9: Example of a bed file, that will be used in this work, related to the H1 cell line and the histone modification H3K4me1. The first column indicates the chromosome (chromosome 15), the second column, the initial position, and the third column, the final position of the DNA fragment contacting the modification histone, and the last column, shows the study (USSC), cell line (H1), histone (H3K4me1) and specificity (LL228). Image taken from MATLAB.

In bedDetail form at other two columns are added thus having the following structure (see example in figure 9):

1) chrom: chromosome scaffold or name.
2) chromStart: first annotation position (starting at 0)

3) chromEnd: last position of an annotation.
4) Name: Name of the region and metadata.
5) Strand of DNA: + or -

This type of file is usually organized in 3 levels: Study – Sample – Experiment. In this work we use:

- Study: all studies (BI, UCSC, USCF-UBC and UW).
- Sample: all histone with bed files.
- Experiment: all experiment data with bed files.

2.2.1. Modification of histones

In this project were studied 31 histone modifications as shown in table 2, these histones are distributed by different cell types (adult and fetal) (see example in figure 10). Annex 4 shows additional characteristics of each histone modification.

H2A.Z	H3K18ac
H2AK5ac	H3K23ac
H2AK9ac	H3K23me2
H2BK5ac	H3K27ac
H2BK12ac	H3K27me3
H2BK15ac	H3K36me3
H2BK20ac	H3K56ac
H2BK120ac	H3K79me1
H3K4ac	H3K79me2
H3K4me1	H3T11ph
H3K4me2	H4K5ac
H3K4me3	H4K8ac
H3K9ac	H4K12ac
H3K9me1	H4K20me1
H3K9me3	H4K91ac
H3K14ac	

Table 2: Histone modifications studied in this project. Two examples: H3K4me1: Histone H3 monomethylated at lysine 4; H3K27ac: Histone H3 acetylated at lysine 27.

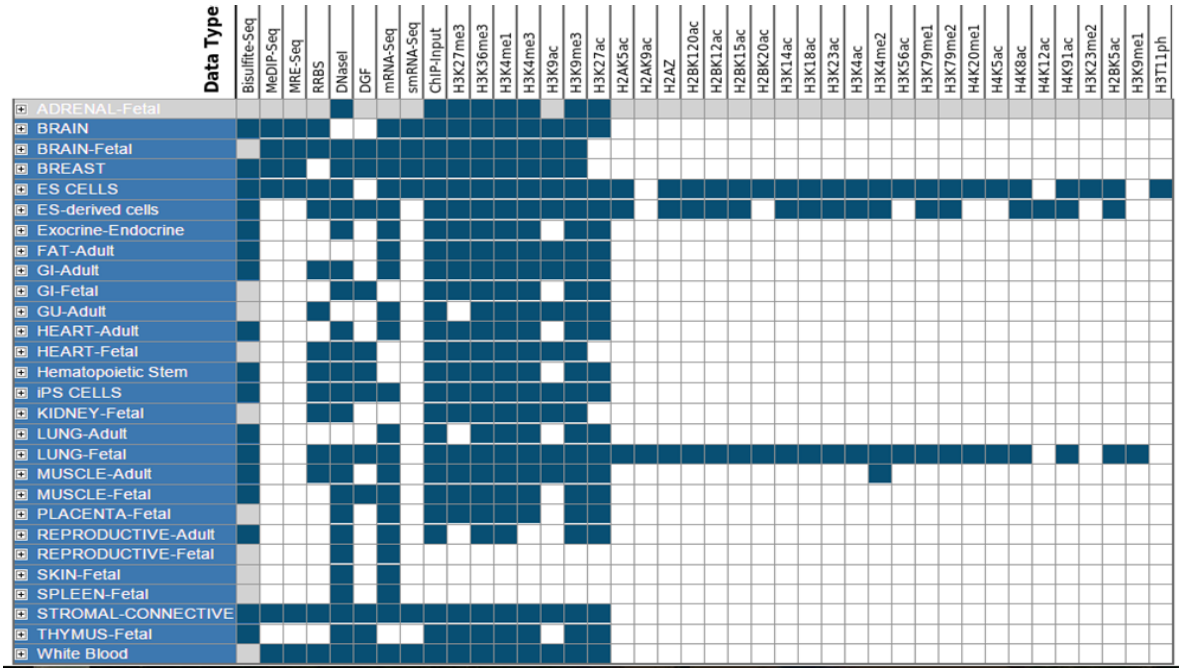


Figure 10: This project used different kinds of histone modifications, depending on the tissue, and the cell type. For example: The brain is represented by 7 histone modifications (H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3 and H3K27ac). It also shows the Bisulfite-Seq, MeDIP-seq and MRE-Seq are three methods for assay the sequencing DNA. RRBS is the reduced representation bisulfite sequencing. The chromatin accessibility can be assayed by sequencing DNase I. RNA expression can be assayed by sequencing mRNAs (mRNA-Seq) or size-selected small RNA fractions to high depths (smRNA-Seq). And finally, Histone modifications can be assayed by sequencing DNA enriched by chromatin immunoprecipitation with modification-specific histone antibodies (ChIP-seq). The histone modifications are experiments, or whatever it is horizontal. And what is vertical are the samples. Imagen taken from the Roadmap Epigenomics (41).

3. Analysis of epigenomes

The main objective of this work is to study the epigenetics signals in the human genome, meaning, whether there is dependence between the DNA sequence and the occurrence of epigenomic marking. The present study employed different statistical methodologies, namely, statistical tests, effect size measures, residue analysis and hierarchical classification. With these methodologies, we compared genomic contexts of epigenetic marking among chromosomes and among different histone modifications, complementing the analysis with a control scenario, with no markings and factoring the C+G content.

3.1. Software tools

In order to analyse the epigenomic data and to perform statistical analysis, we used the R program, a free software for statistical computing and graphing, that have programming language named "S", that is simple and effective. This program obtains data manipulation, calculation. It contains tools and graphics for data analysis and their display (42).

In order to accomplish the desired statistical analysis, we had to install some packages, such as Biostrings gives an evolved environment for efficient flow management and analysis in R It has many utilities, such as many speed and memory effective and string matching algorithms for fast manipulation of large sets of sequences (43). Through objects and functions provided by Biostrings originates many other sequence analysis packages. Also, this package allows to exclude undesired regions through mask XString objects, so that these masks can be turned on and off any time. The XSTring also allows to represent the different types of biosequence strings in subclasses: BString, RNAString, DNAString and AAString. Biostrings contains functions for performing many of the basic sequence transformation and manipulation routines and it makes also possible performing multiple sequence alignments (43).

Another package used was WriteXLS, which serves to create an Excel file with data obtained from the R program (44). Package XLSX, was also used with the objective of working with the Excel file data in R program (45).

Another important package used was the “GenomicRanges package that defines general purpose containers for storing and manipulating genomic intervals and variables

defined along a genome" (46). This package serves as the foundation for representing genomic locations within the Bioconductor project, giving tools for the analysis and comprehension of high-throughput genomic data. "The Bioconductor hierarchy, is built upon the IRanges (infrastructure) function and provides support for the BSgenome (infrastructure) and GenomicFeatures (infrastructure) functions, and many other functions". The IRanges package introduces three classes (GRanges, GPos, and GRangesList), which are used to represent genomic ranges, genomic positions, and groups of genomic ranges, respectively (46). We used the "GRanges class that represents a collection of genomic ranges where each have a single start and end location in the genome". Therefore, it may be used to store the location of genomic features, for example, contiguous binding sites, transcripts, and exons (46).

In order to analyse the genomic context of epigenetic marking, we used some R scripts in the program R. One of these altered columns that were relevant of the analysis, because the database provided additional information that was not needed. Another R script was created to obtain the frequency of oligonucleotides, i.e. single nucleotides, dinucleotides, trinucleotides and tetranucleotides. Wherein, first we obtained the frequency of oligonucleotides for fragment, where each fragment is a regional a chromosome with start and an end position. This function counted the frequency of oligonucleotides contained in that space (between the initial position and the end position). To obtain the full frequency of oligonucleotides for the chromosome 1, for example, the total of the fragments present in this chromosome was summed.

Besides these two R scripts mentioned here, we used others, which are described in detail annex 5.

3.2. Statistical procedures

In this work we have used t-test, Cohen's d, chi-square test, Cramer's V, residual analyses and hierarchical clustering, to explore the possibility of dependence between DNA contexts and epigenetic marking occurrence.

3.2.1. Effect size

The effect size is a measure independent of the sample size that quantifies the force with which a phenomenon of interest occurs, that is, it is a descriptive statistic that serves as a complement to the test of statistical significance (47). The Cohen's d is one

effect size measure, which can be used to complement t-test results for example, by evaluating the mean difference between two groups. The Cohen's d measure is arbitrarily classified. One common classification of Cohen's d effect size is:

(d = 0.8) → large

(d = 0.5) → moderate

(d = 0.2) → small

So, in these statistical analyses we calculated the Cohen's d, using the equation:

$$d = \frac{\overline{X_1} - \overline{X_2}}{s_{pooled}}$$

Where:

$$S_{pooled} = \sqrt{\frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

$\overline{X_1}$ e $\overline{X_2}$- The mean of the group 1 and group 2, respectively.

S_1^2 e S_2^2- The sample variances of the group 1 and group 2, respectively.

n_1 e n_2- The sample sizes in the group 1 and group 2, respectively.

The concept of contingency table will also be required from now on. This is a type of table, in a matrix format, that displays the frequency distribution of two variables and can be represented generically by as in table 3:

		Y			
		y_j	...	y_c	
X	x_i	n_{rc}	...	n_{ic}	$n_{i.}$
	⋮				⋮
	x_r	n_{rj}	...	n_{rc}	$n_{r.}$
		$n_{.j}$	...	$n_{.c}$	n_{ij}

Table 3: The contingency table. n_{ij} – Total of observations; $n_{i.}$ – The total observations in the category i variable row; $n_{.j}$ – The total of observations in the category j variable column.

Cramer's V is another effect size measurement, which is associated with the chi-square test. Alternatively one can also use phi coefficient. The Cramer's V describes the combination of intensity in the sample and takes values between 0 and 1. The 0 value corresponds to the absence of association between variables, near zero values correspond to weak association and values closer to 1 values correspond to strong association (48). As with Cohen's d the classification of the V effect force is also subjective, so one frequent classification associated with biological sciences is as follows:

(V ≈ 0.1) → Weak association

(V ≈ 0.3) → Moderate association

(V ≈ 0.5) → Strong association

In this work, we considered that the Cramer's V assumed a negligible value if lower than 0.001, and for values bigger than 0.001 will not be considered negligible.

This measure of association is calculated by:

$$V = \sqrt{\frac{X^2}{n(K-1)}}$$

n- Total of observations

K- The minimum {c, r}

On the other hand, the phi coefficient can be calculated by:

$$\Phi = \sqrt{\frac{X^2}{n}}$$

n- Total of observations

It is important to highlight that small and very small effects may be significant, because just for this that the sample size is large enough. Herein, in order to make a more complete analysis, the results of statistical tests will always be supplemented with the values of the corresponding measures of effect size.

3.2.2. Statistical test

For statistical testing it is important to define the null hypothesis and in turn, the alternative hypothesis. The null hypothesis (H_0) must be formulated in order to be rejected, if an effect exists. Otherwise, the test won't add nothing about the characterization of the population (48,49). Conversely, the alternative hypothesis (H_1) is associated with the theory to demonstrate (48,49).

In addition to defining H_0 and H_1, to perform the hypothesis test it is also necessary to establish a level of significance (α). After words, the decision of the hypothesis test is based on the p value. The hypothesis H_0 is rejected if the p value is less than the significance level, and the most used significance levels are 5%, 1% and 10% (48,49).

We used the t-test with the aim of comparing two independent samples. This test compared the average of a quantitative variable in two different groups of subjects and is unaware of their variances. The conditions for its use are:

i. Two independent aleatory samples;
ii. Two populations normally distributed with unknown variance.

The t-test is calculated by:

$$T = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

n_1 e n_2- The sample size of the group 1 and group 2, respectively.

The chi-square test serves to test if the distributions of two or more unrelated samples differ significantly. However, this test can also be used to test the independence between two variables or the homogeneity between groups (48,49).

The conditions for the use of this test are:

i. Exclusively for nominal and ordinal variables;
ii. Preferably for large samples, >30;
iii. Independent observations;
iv. It does not apply if 20% of the expected values are below 5.

We can calculate the chi-square value, using the formula:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

O- Observed values

E- Expected values

High statistical values lead to the rejection of the H_0 (the hypothesis of independence and equality in the distribution of different groups), for the benefit of H_1.

3.2.3. Residual analysis

With the purpose of identifying the categories responsible for a significant value of the statistical chi-square, test a process that involves the analysis of normalized residuals is used. The normalized residuals are calculated as follows:

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

n_{ij} - Total observations in contingency lines and columns.

e_{ij} - Expected frequencies.

To find the expected value, the following formula is used:

$$e_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

$n_{i.}$ – The total observations in the category i variable row.

$n_{.j}$ – The total of observations in the category j variable column.

An estimate of the variance of r_{ij} can be obtained using:

$$\widehat{v_{ij}} = \left(1 - \frac{n_{i.}}{n}\right)\left(1 - \frac{n_{.j}}{n}\right)$$

Then, for each cell of the contingency table, we can calculate the standardized residual (d_{ij}) using:

$$d_{ij} = \frac{r_{ij}}{\sqrt{\widehat{v_{ij}}}}$$

To interpret a residual analysis in the context of contingency tables, a heatmap can be created in R program, which is a two-dimensional representation of data in which values are represented by colours to indicate the level of intensity. Normally, darker colours indicate lower effect size, while brighter colours indicate higher effect size.

3.2.4. Clustering methods

The clustering methods that were used are hierarchical methods that result in partition hierarchies. Hierarchical methods can be further subdivided into agglomerative and divisive methods (50). The application of agglomerative methods starts with a partition with as many parts as the number of different individuals that will because grouped, while the divisive methods considers as a starting point a partition with a single group, a single set including all individuals that will be divided in hierarchical groups (50).

In this work we have used a process of grouping or aggregation that determines similarities between individuals using a similarity measure and performs the grouping of subjects in groups via an aggregation policy so that they can be represented by a dendrogram (50). The dendrogram has the advantage of facilitating viewing of the clustering process in its various phases from separate individuals to a single a unified group. In fact, the dendrogram is a graphical representation of a partition hierarchy, wherein each group is called a cluster.

We have chosen agglomerative methods, since these are the most used and most reported in the literature. One of the reasons for this is that it needs less computational effort for the same kind of analysis. The clustering algorithms group data according to similarity indices.

For agglomerative methods there are several aggregation criteria, and what distinguishes them is the way of setting distance between groups (50). Examples of aggregation criteria are single linkage, complete linkage and average.

- Single linkage: is one of the simple algorithms for hierarchical clustering and used the nearest neighbor technique, i.e. the distance between two clusters is determined by the distance between the closest groups. However,

using this algorithm, in the presence of distant groups located between two clusters, it forms a bridge and force an inappropriate combination of these clusters – the so called chaining effect (50).

- Complete linkage: it is known as the technique of the farthest neighbor, since it determines the distance between two clusters according to the longest distance between a pair of groups, a mode most likely to identify clusters that are less extended (50).
- Average linkage: Under this option, the distance between two clusters is defined as the average of the distances between all sample pairs on each cluster (50).

One cannot say that there is a criterion of aggregation that is better than the others, that is, in practice, what is done is to use multiple criteria and to compare the results. If the results obtained with different criteria are concordant the end result can be considered "more credible".

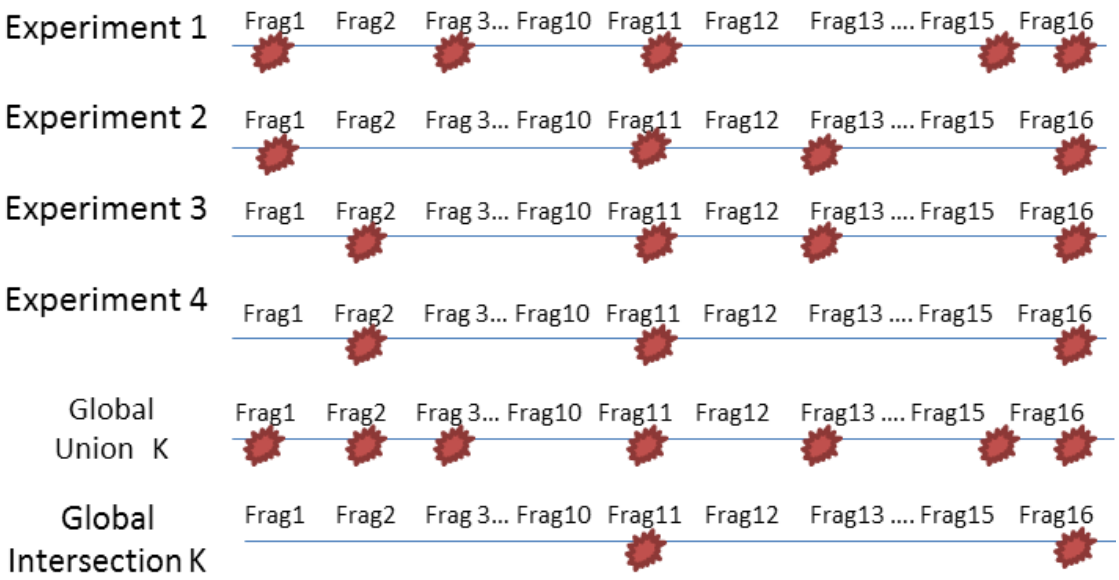
Finally, when data are organized in attribute-value tables, their attributes can take different types: binary, discrete or continuous. Binary attributes assume only two values, for example yes and no or 1 and 0, indicating the presence or absence of a particular feature. Discrete attributes have often a finite and small set of possible values. Continuous attributes can take any real value in one of a pre-defined range, which will be our case. The interpretation of dendograms varies with the methods of similarity and different grouping techniques used. For example, complete linkage clustering should produce higher distance values between large clusters, while a dendrogram generated by the single link algorithm tends to have smaller distance values for the same cluster (50).

3.3. Procedure used for epigenomes analysis

3.3.1. Epigenomic regions manipulation

In order to analyse all epigenomes, we decided to use both the union and the intersection. The union is when two or more sets are put together, establishing a common relationship between its elements. The intersection is when two or more elements are common to related sets. In our analysis we have:

- Global union and global intersection will be used for the analysis of whole genome, focusing on the context epigenetic marking on chromosomes, as exemplified in figure 11:



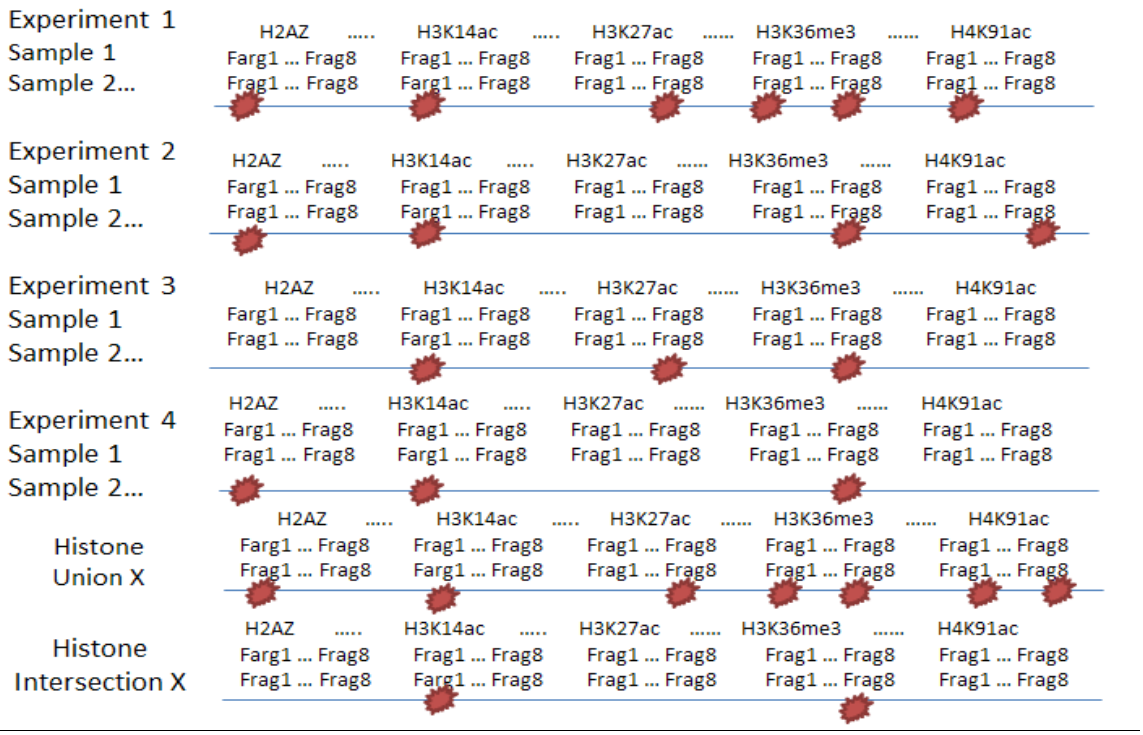
Location of an epigenetic mark.

K - Total of experiments used.

Figure 11: Global union and global intersection will be used for the analysis of whole genome. The fragments are pieces of DNA that can be there this epigenetic marking.

- Histone union and histone intersection will be used for histone analysis. We have used data from all histone modifications globally, to create the union and the intersection sets, as exemplified in figure 12:

Each experiment can have multiple samples and we know where each epigenetic marking occurs. Thus, we intend to study the context of these occurrences of epigenetic marking in these samples. Limiting the data will give us a picture of all positions (union) where a modification has already been found, while the intersection will only highlight the positions where all samples have a modification, i.e. those begging the showiest signal for modification.



Location of an epigenetic mark.

X – Total of sample of each experiment.

Figure 12: Histone union and histone intersection will be used for histone analysis. The fragments are pieces of DNA that can be there this epigenetic marking.

3.3.2. Non-marked regions – Control

All regions without epigenetic marking will be used as control, consisting of 43916 fragments. The control context is represented by 19369407 nucleotides, has 5619417 A nucleotides; 5625109 T nucleotides; 4063926 C nucleotides; and 4060955 G nucleotides.

This control was built to be able to compose the difference between marked and un-marked regions, from the point of view of nucleotide contexts.

3.3.3. Epigenomic regions features

After obtaining the bed files about histone modifications, we calculating the union and intersection of data for each histone (table 4). We were able to obtain the number of fragments, either for the union or for the intersection. Then, we went to see the each size histone by the results of nucleotides each. Regarding the total number of epigenomes of each histone is present in Epigenomes Roadmap.

Histone modification	Total number of epigenomes used	Number of fragments of intersection	Size of intersection	Number of fragments of union	Size of union
H2AZ	9	943538	87595468	3370563	5662528428
H2AK5ac	14	1099193	19392529	5118048	5203388932
H2AK9ac	2	7540982	187681697	25196481	1252156479
H2BK5ac	14	1367558	187681697	6108154	5168729322
H2BK12ac	12	756268	15073138	5374386	4883438272
H2BK15ac	12	755162	43077505	2882531	5271808939
H2BK20ac	5	2591040	250722582	5958391	4684430742
H2BK120ac	13	119170	6646099	4010072	5035054959
H3K4ac	13	5125539	9765152	3870935	5075999683
H3K4me1	219	541	20425	263132	5641226891
H3K4me2	15	680423	36200326	2985055	5344793021
H3K4me3	222	35942	3657742	269427	5640204628
H3K9ac	92	116596	3090979	397897	5610818048
H3K9me1	2	5193073	994475518	7398570	2800075700
H3K9me3	222	12220	570572	232374	5651302475
H3K14ac	11	1590559	38913268	4394035	4976720387
H3K18ac	14	572023	14272972	4851227	4885264835
H3K23ac	13	202129	2554983	6304441	5085018511
H3K23me2	4	2514007	86889503	9726522	4137734719
H3K27ac	155	16679	1122336	305110	5629002434
H3K27me3	220	3303	150566	240131	5647724605
H3K36me3	224	5057	159604	247963	5645795013
H3K56ac	6	704098	55084013	6307120	4694914176
H3K79me1	16	524723	8958925	5864587	5149040554
H3K79me2	9	1499388	86792699	5036068	5080791287
H3T11ph	1	22475850	919530849	22475850	919530849
H4K5ac	6	1502077	4746029261	5592228	4746029261
H4K8ac	14	135657	3066909	5791131	4644321334
H4K12ac	2	4550850	913428784	6573836	2531942517
H4K20me1	5	665657	59172393	7343572	4201096750
H4K91ac	9	748451	17829434	5418264	4604027535

Table 4: Overall analysis of the union and intersection for each histone modification, the columns present the total number of epigenomes, size of reunion, size of intersection and number of fragments.

As would be expected, when there is only one epigenome the union and intersection will be the same (ex: H3T11ph). When we increase the number of epigenomes, we have greater union values composed to a smaller intersection (ex: H2AK9ac).

In relation to the chromosomes, we also made a global analysis getting the intersection, union and control (table 5). The size was obtained by nucleotides from all fragments present in each chromosome.

	Global					
	Intersection		Union		Control	
Parameters	Number of fragments	Size	Number of fragments	Size	Number of fragments	Size
Chr1	16	1989	14482	445209149	3759	2020432
Chr10	4	131	10822	258749690	2965	1439029
Chr11	0	0	5181	261238234	1005	311331
Chr12	0	0	4508	260263443	832	180427
Chr13	0	0	3159	190662634	571	139927
Chr14	0	0	4099	175892351	867	173738
Chr15	0	0	9615	159858419	2837	1302962
Chr16	0	0	8086	155093837	2101	969623
Chr17	0	0	4450	154057373	1088	574219
Chr18	7	322	2618	148871408	522	120858
Chr19	19	653	2403	110732367	601	338352
Chr2	3	73	11226	474279089	2164	619698
Chr20	0	0	1628	118751369	315	67217
Chr21	0	0	1598	70001825	270	47194
Chr22	0	0	3919	68904437	966	262435
Chr3	0	0	6705	388557523	1264	262379
Chr4	0	0	7730	373806895	1441	465539
Chr5	0	0	10105	351525659	2473	1489023
Chr6	2	180	6517	333579704	1285	354256
Chr7	3	242	9429	308401930	2077	764556
Chr8	0	0	7377	283606922	1770	769314
Chr9	0	0	25186	230842567	7412	3461339
ChrX	0	0	12853	298904834	2486	1147443
ChrY	0	0	12594	40736769	2845	2088116

Table 5: Global analysis showing the intersection, union and control, number of fragments and size.

As can be seen the intersection only has data for 7 chromosomes, unlike the union and control where all chromosomes are represented (24 chromosomes). Again, the number of fragments is smaller for the intersection forger for the union.

3.4. Genomic context analysis

3.4.1. Content C+G – Control vs Union

One characteristic of DNA is its C+G content. The C+G content various greatly between organisms and within the gene. The of a single organism, especially among mammals. It is also know that the human genome has regions of high C+G content, alternating with regions of low C+G content.

To move away the hypothesis that the C+G contents could to evaluate the association between C+G content and the occurrence of epigenetic marking, we applied the t-test to test if:

H_0: C+G content of average union= C+G content of average control

H_1: C+G content of average union≠ C+G content of average control

We obtained the following results (table 6):

t-test	df	p-value	Cohen's d
7.33	229261	<0.001	0.039

Table 6: T-test values and Cohen's d. df matches degrees of freedom. p-value matches p-value

Thus allowing to reject the H_0 hypothesis. We also built a box plot (see figure 13), which is a graphical representation frequently used to compare multiple data sets. With this graph we can see how the data are distributed, namely, the highest or lowest concentrations, symmetry and occurrence of outlier values.

For this representation, one begins by collecting 5 types of information from each sample: the two extremes (minimum and maximum), the median and the 1st and 3rd quartiles. Then, these values are plotted for each sample of values.

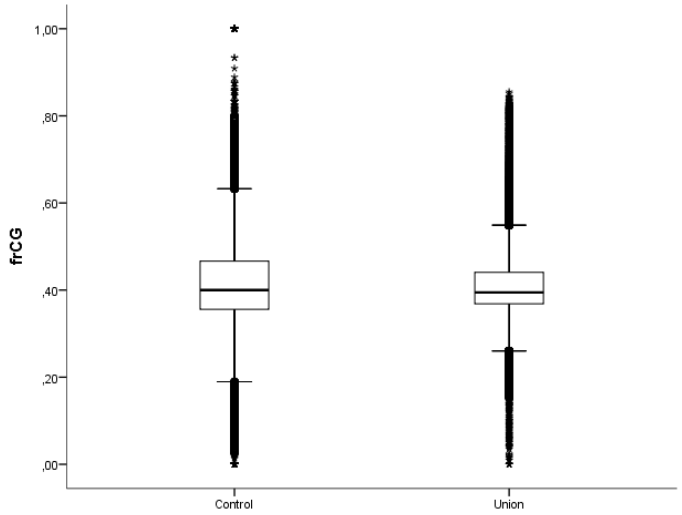


Figure 13: Boxplot of C+G content for control and union. The union has outliers, like the control. This means that market fragments are sometimes very rich in C and G, something that happen in un-market ones.

As we can see in figure 13, the union set and the control set have many higher outliers and have some lower outliers.

As mentioned, Cohen's d is a measure of the force with which a phenomenon of interest occurs, serving that way in complement to the p-value. Here, we obtained the following result: d = 0.039. Thus, by classification of Cohen's d, we conclude that the size effect of C+G content of our analysis is small.

3.4.2. Global analysis of genome

We compared the total of fragments of the global union with those of the control, to understand if the context of epigenetic marking is equal and thus there is no specificity for epigenetic marking at the level of DNA sequence. The same held for global intersection. To do this, we applied the chi-square test and the Cramer's V, for nucleotides, dinucleotides, trinucleotides and tetranucleotides (table 7).

	Union vs Control				Intersection vs Control			
Parameters	X^2	df	p-value	Cramer's V	X^2	df	p-value	Cramer's V
Nucleotide	8969.3	3	*	0.0012	15.9	3	0.0012	0.0009
Dinucleotide	29912	15	*	0.0022	202.47	15	*	0.0032
Trinucleotide	54176	63	*	0.0030	645.48	63	*	0.0057
Tetranucleotide	91899	255	*	0.0040	2208.8	255	*	0.0107

Table 7: Chi-square test values for comparing the total of fragments of the global union with the control. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p-value; Cramer's V matches measure of association Cramer's V. * p-value is <0.001.

The hypotheses tested were:

H_0: The epigenetic marking context in marked regions is equal to context in control unmarked ones.

H_1: The epigenetic marking context in marked regions is not equal to context in control unmarked ones.

Looking at table 7, we can see that the epigenetic marking contexts are not equal since we rejected the H_0 hypothesis. From values of Cramer's V, we see that the union have a stronger heterogeneity effect. However, even with small values for the Cramer's V we proceeded to the analysis of residual.

Using the heatmap function of the R program to visualize the analysis of residual of values from table 7, we conducted heatmaps for nucleotides, dinucleotides, trinucleotides and tetranucleotides. However, we will only show the union (see in figure 14) and the intersection (see in figure 15) for dinucleotides. The rest of non-neglectable heatmaps are in annex 8.

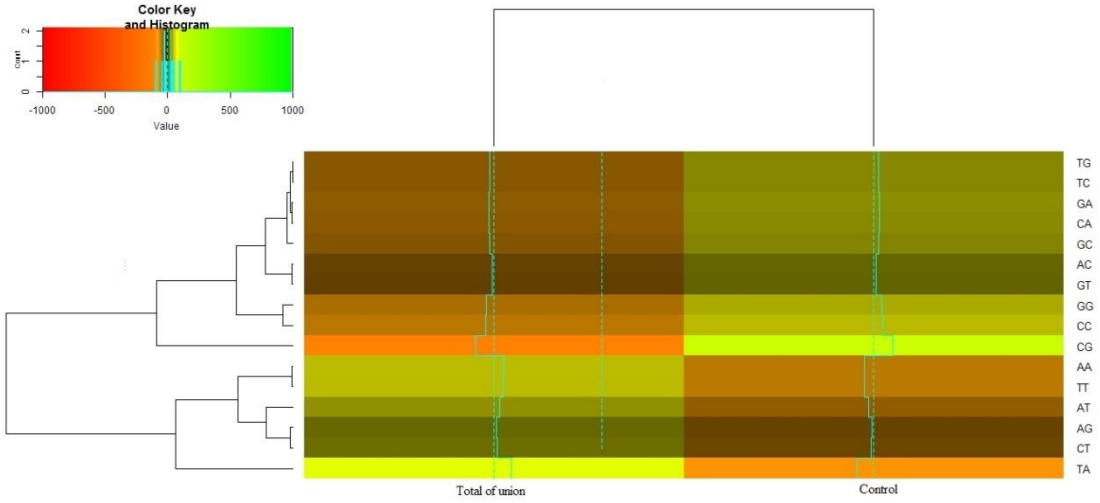


Figure 14: Heatmap which corresponds to comparison for dinucleotides union. The preference for both types of fragments is opposite.

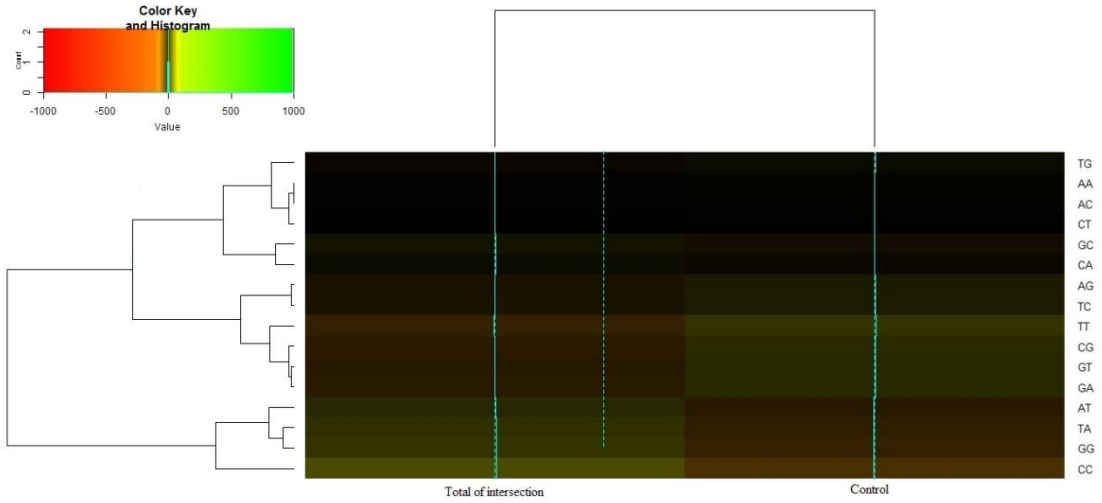


Figure 15: Heatmap which corresponds to comparison with the control for dinucleotides intersection. We can again see that the preference for both types of fragments is opposite.

In both cases, marked and unmarked regions show preference for opposite contexts. However, in the case of union marked regions prefer the dinucleotides TA, while the control prefers CG. In the case of intersection marked regions prefer dinucleotide CC and unmarked ones TT.

3.4.3. Global analysis of histones

In order to test if the context of epigenetic marking is homogeneous among histone modifications. We applied the chi-square test and the Cramer's V for histone union and histone intersection again for nucleotides, dinucleotides, trinucleotides and tetranucleotides (table 8).

	Histone union				Histone intersection			
Parameters	X^2	df	p-value	Cramer's V	X^2	df	p-value	Cramer's V
Nucleotide	78802000	90	*	0.013	56302000	90	*	0.0466
Dinucleotide	173030000	450	*	0.0089	132060000	450	*	0.0320
Trinucleotide	255120000	1890	*	0.0076	200350000	1890	*	0.0280
Tetranucleotide	335490000	7650	*	0.0087	335490000	7650	*	0.0087

Table 8:Chi-square test value for the histone union and histone intersection of the 31 histone modifications. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p value; Cramer's V matches measure of association Cramer's V. *p-value is <0.001.

Looking at the hypotheses of this test we have:

H_0: The epigenetic marking context is homogeneous among histone modifications.

H_1: The epigenetic marking context is not homogeneous among histone modifications.

Through the values obtained for the chi-square test, we concluded that epigenetic marking context is not homogeneous among histones, either for histone union or to the histone intersection. With regard to the values of Cramer's V, there is a weak heterogeneity effect either the histone union or the histone intersection, but the histone intersection has a stronger heterogeneity effect.

Despite the values of Cramer's V being low, we also performed the residuals analysis. Although this was done for nucleotides, dinucleotides, trinucleotides and tetranucleotides, we only present the heatmap of the histone union (see in figure 16) and the histone intersection for single nucleotides (see in figure 17), because the remaining had the same profile. The remaining heatmaps are presented in annex 8.

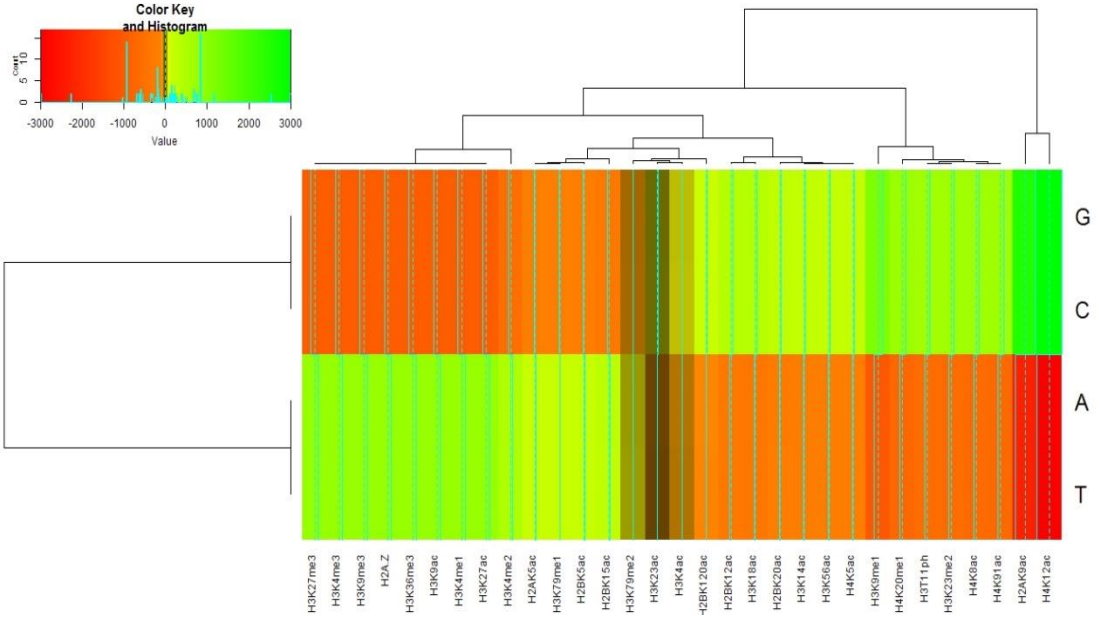


Figure 16: Heatmap of nucleotides histone union of all histone modifications. There are two histone modifications that stand out, the H4K12ac and H2AK9ac, which have preference for the nucleotides G and C. All other histones has a more homogeneous preference and show an environment in As and Ts.

Through the analysis of figure 16, there are two histone modifications that show a different pattern in relation to the preference of the DNA nucleotide content, these being H4K12ac and H2AK9ac who have preference for nucleotides G and C. There were four clusters with homogeneous behaviors, one of which being, with preference for A and T

nucleotides, formed mainly by histone H3 modifications and methylations plus two acetylations and histone H2AZ. Another cluster consists of almost all histone H4 modifications, having also three histone H3 modifications, with preference for G and C nucleotides. The third cluster is formed by H4K12ac and H2AK9ac and shows preference for nucleotides G and C. And the last cluster is formed by the remaining histone H3 and H2 modifications, mainly acetylation, and having preference for nucleotides G and C or T and A. Thus, we conclude that the groups may be formed by the type of histone (H2, H3 and H4), as well as by the type of modification (acetylation or methylation). In a more general form, we can see groups formed by transcription activating histone modifications (normally acetylations) while other groups include mainly modifications of histones that inactivate transcription (normally methylations).

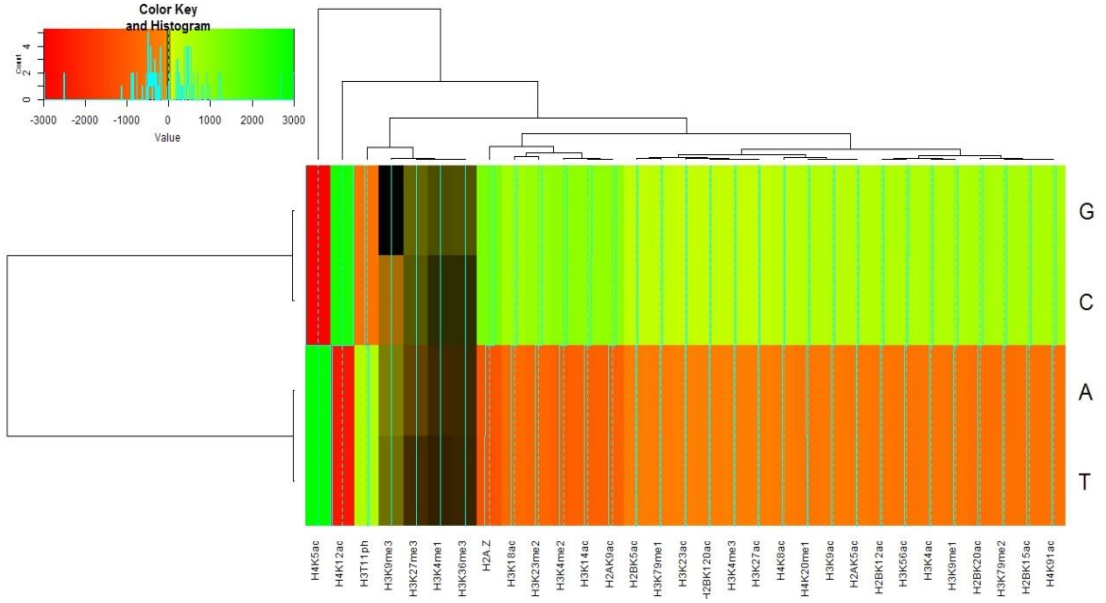


Figure 17: Heatmap of nucleotides histone intersection of all histone modifications. There are two histone modifications that stand out, the H4K5ac that has preference for the content of nucleotides A and T and H4K12ac that has preference for the content of nucleotides G and C. All other histones has a more homogeneous preference and show an environment in As and Ts.

Relative to intersection, figure 17, we again have two histones that stand out in relation to the preference of the content, histone H4K5ac has preference for the content of nucleotides A and T; and histone H4K12ac has preference for the content of nucleotides G and C, as in the histone union dataset. All other histone had a more homogeneous behavior. Therefore, it is possible to identify a cluster formed only by H4K5ac; another one formed by H4K12ac; and a third cluster, which is subdivided into several sub-clusters formed by other histone modifications, most of which preferring G

and C rich contexts. Here it was not possible to identify groups by neither the type of histone nor the histone modification, like in the case of union.

We can conclude for this analysis that histone modifications can be associated to specific contexts, enriched in either A and T or G and C.

3.4.4. Chromosomes analysis

In this analysis, we wanted to evaluate if the genomic context associated with the occurrence of epigenetic marking is homogeneous, among chromosomes. For this, we applied the chi-square test and the Cramer's V value (table 9), on the global union and global intersection datasets, for nucleotide, dinucleotide, trinucleotide and tetranucleotide contexts.

	Global union					Control				
Parameters	X^2	df	p-value	Cramer's V	n	X^2	df	p-value	Cramer's V	n
Nucleotide	10063000	69	*	0.0243	11325056856	57981	69	0.001	0.0315	38738814
Dinucleotide	22290000	345	*	0.0161	11324685040	139530	345	*	0.0219	38651860
Trinucleotide	34156000	1449	*	0.0161	11324313240	232510	1449	*	0.0228	38565276
Tetranucleotide	46218000	5865	*	0.0188	11323941448	363210	5865	*	0.0286	38479184

Table 9: Chi-square test to evaluate the homogeneity between epigenetic marking regions and non-epigenetic marking regions for the global union datasets. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p-value; Cramer's V matches measure of association Cramer's V; n matches the sample size. *p-value is <0.001

The test was conducted for the following hypotheses:

H_0: The nucleotide context having epigenetic marking is homogeneous among chromosomes.

H_1: The nucleotide context having epigenetic marking is not homogeneous among chromosomes.

Through this test, we concluded that epigenetic marking is not homogenous among chromosomes. Regarding the Cramer's V value, the heterogeneity effect is weak in both global union and unmarked control, but the effect in the control showed a stronger heterogeneity effect than the global union. Although Cramer's V values are very low, there is indeed a significant but weak heterogeneity effect.

Therefore, we performed a residual analysis to better understand where such small differences could be located.

Although not presented here, the same procedure was carried out for the global interception. However, that dataset was limited to only 7 chromosomes, yielding less relevant results.

Therefore, we have built the heatmap with the residuals for the global union only (see in figure 18). We only show here the results for nucleotide contexts, while those for dinucleotides, trinucleotides and tetranucleotides are in annex 8.

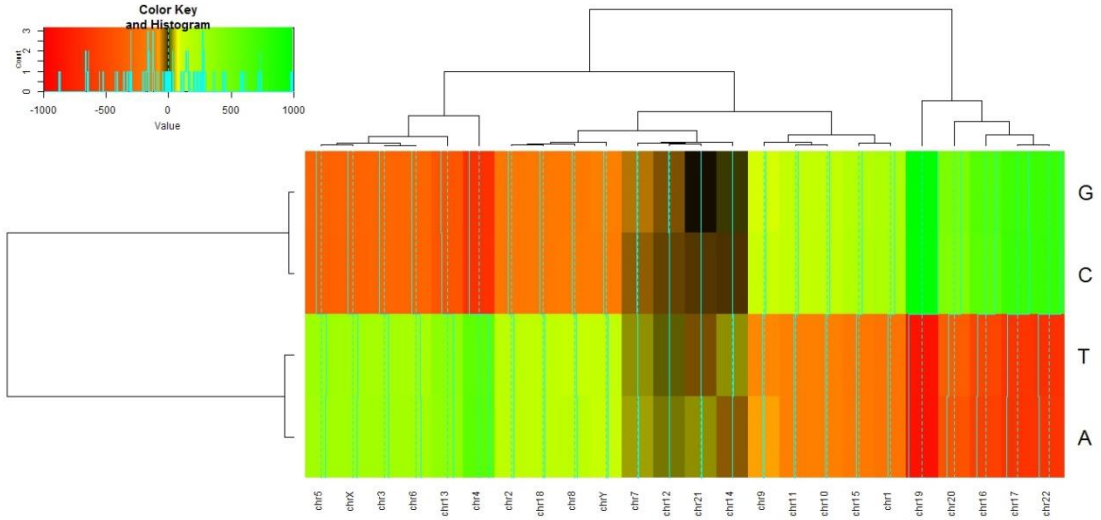


Figure 18: Heatmap of global union for single nucleotides. Three clusters were formed, two of which have a preference for the nucleotides A and T or G and C, and another cluster much more homogenous.

Analysing the heatmap of figure 18, we can conclude that there are contexts identical profiles in the various chromosomes, showing preference for nucleotides A and T or C and G. There are 3 groups where there is homogeneity among chromosomes: one formed by chromosomes 19, 20, 16, 17 and 22 which have preference for nucleotides G and C; another formed by chromosomes 5, X, 3, 6, 13 and 4, with preference for nucleotides T and A; and, finally, another cluster formed by the remaining chromosomes that have a homogeneous preference for all nucleotides. Still other chromosomes showed isolate preference for nucleotides A and T or G and C.

We conclude that different chromosomes have different preference for contexts enriched in either A and T or G and C. It also was possible to identify specific genomic contexts in some chromosomes through the heatmaps of trinucleotides (table 10):

Chromosome	Identify specific genomic contexts in some chromosomes
Chr3 Chr4 Chr5 Chr6 Chr13 ChrX	TAT and ATA
Chr15 Chr16 Chr17 Chr20 Chr22	GGG, CCC, GCC and GGC
Chr19	GGG, CCC, GCC, GGC, CCG, CGG, GCG and CGC

Table 10: Identify specific genomic contexts in some chromosomes.

We carried out a more detailed analysis, with the intention of knowing which were the chromosomes that stand out relatively to the data for each histone modifications, i.e., to evaluate the homogeneity of each histone modifications, chromosome by chromosome. To do this, again we applied the chi-square test and the Cramer's V value to the global union dataset (table 11).

Chromosome		X^2	Df	p-value	Cramer's V
Chr1	Nucleotide	5541900	90	*	0.01258094
	Dinucleotide	12210000	450	*	0.008355758
	Trinucleotide	17976000	1890	*	0.007172909
	Tetranucleotide	23507000	7650	*	0.008168113
Chr10	Nucleotide	3174700	90	*	0.01251314
	Dinucleotide	7016300	450	*	0.008323797
	Trinucleotide	10358000	1890	*	0.00715518
	Tetranucleotide	13605000	7650	*	0.008165683
Chr11	Nucleotide	3715100	90	*	0.01347588
	Dinucleotide	8161500	450	*	0.008937444
	Trinucleotide	12022000	1890	*	0.0076742
	Tetranucleotide	15773000	7650	*	0.008753228

Chr12	Nucleotide	3235000	90	*	0.01258799
	Dinucleotide	7101900	450	*	0.008345724
	Trinucleotide	10488000	1890	*	0.007175305
	Tetranucleotide	13787000	7650	*	0.00819161
Chr13	Nucleotide	3009100	90	*	0.01443056
	Dinucleotide	6590500	450	*	0.009556667
	Trinucleotide	9755500	1890	*	0.008226674
	Tetranucleotide	12925000	7650	*	0.009419966
Chr14	Nucleotide	2412200	90	*	0.01329056
	Dinucleotide	5287000	450	*	0.00880438
	Trinucleotide	7790400	1890	*	0.007561428
	Tetranucleotide	10241000	7650	*	0.008631028
Chr15	Nucleotide	1661400	90	*	0.01150411
	Dinucleotide	3682000	450	*	0.007663108
	Trinucleotide	5437600	1890	*	0.006588455
	Tetranucleotide	7122800	7650	*	0.007508651
Chr16	Nucleotide	1391600	90	*	0.01061339
	Dinucleotide	3130200	450	*	0.007122217
	Trinucleotide	4638200	1890	*	0.006133536
	Tetranucleotide	6118800	7650	*	0.007019477
Chr17	Nucleotide	1305500	90	*	0.01021069
	Dinucleotide	2939600	450	*	0.006855525
	Trinucleotide	4350500	1890	*	0.005900137
	Tetranucleotide	5704200	7650	*	0.006733217
Chr18	Nucleotide	2035600	90	*	0.01330172
	Dinucleotide	4482200	450	*	0.008832395
	Trinucleotide	6632000	1890	*	0.007601348
	Tetranucleotide	8755600	7650	*	0.008693676
Chr19	Nucleotide	760310	90	*	0.009184634

	Dinucleotide	1753800	450	*	0.006241502
	Trinucleotide	2621000	1890	*	0.005397956
	Tetranucleotide	3536500	7650	*	0.006251441
Chr2	Nucleotide	6482600	90	*	0.01329926
	Dinucleotide	14234000	450	*	0.008818181
	Trinucleotide	21020000	1890	*	0.00758159
	Tetranucleotide	27659000	7650	*	0.008656126
Chr20	Nucleotide	1074300	90	*	0.01054775
	Dinucleotide	2422900	450	*	0.007087515
	Trinucleotide	3588100	1890	*	0.006101751
	Tetranucleotide	4705200	7650	*	0.006965709
Chr21	Nucleotide	1267000	90	*	0.01536944
	Dinucleotide	2779500	450	*	0.01018642
	Trinucleotide	4086200	1890	*	0.008738524
	Tetranucleotide	5387900	7650	*	0.009987341
Chr22	Nucleotide	400980	90	*	0.008454151
	Dinucleotide	946190	450	*	0.005810546
	Trinucleotide	1426000	1890	*	0.005046394
	Tetranucleotide	1904400	7650	*	0.005813841
Chr3	Nucleotide	5199900	90	*	0.01314445
	Dinucleotide	11400000	450	*	0.008708961
	Trinucleotide	16868000	1890	*	0.007495008
	Tetranucleotide	22276000	7650	*	0.008573314
Chr4	Nucleotide	5640000	90	*	0.01417404
	Dinucleotide	12321000	450	*	0.00937503
	Trinucleotide	18278000	1890	*	0.008079123
	Tetranucleotide	24295000	7650	*	0.009263419
Chr5	Nucleotide	4899400	90	*	0.01348501
	Dinucleotide	10751000	450	*	0.008938925
	Trinucleotide	15922000	1890	*	0.007696586

	Tetranucleotide	21071000	7650	*	0.008810834
Chr6	Nucleotide	4500600	90	*	0.0132051
	Dinucleotide	9860000	450	*	0.008746035
	Trinucleotide	14584000	1890	*	0.007525744
	Tetranucleotide	19220000	7650	*	0.008598862
Chr7	Nucleotide	4114300	90	*	0.01317627
	Dinucleotide	9023000	450	*	0.008731506
	Trinucleotide	13299000	1890	*	0.007500061
	Tetranucleotide	17470000	7650	*	0.00855484
Chr8	Nucleotide	3760900	90	*	0.01309428
	Dinucleotide	8268900	450	*	0.008688105
	Trinucleotide	12231000	1890	*	0.007475839
	Tetranucleotide	16117000	7650	*	0.008542491
Chr9	Nucleotide	3111900	90	*	0.01331718
	Dinucleotide	6836800	450	*	0.00883256
	Trinucleotide	10065000	1890	*	0.007582296
	Tetranucleotide	13252000	7650	*	0.00865695
ChrX	Nucleotide	4437700	90	*	0.01453681
	Dinucleotide	9792200	450	*	0.009664377
	Trinucleotide	14648000	1890	*	0.008364436
	Tetranucleotide	19642000	7650	*	0.009605598
ChrY	Nucleotide	585460	90	*	0.01747597
	Dinucleotide	1367000	450	*	0.01195572
	Trinucleotide	2205700	1890	*	0.01075021
	Tetranucleotide	3500200	7650	*	0.01326867

Table 11: Chi-square test value for the global union chromosome by chromosome. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p value; Cramer's V matches measure of association Cramer's V. *p-value is <0.001

The hypotheses test was:

H_0: The epigenetic marking context is homogeneous among histones chromosome by chromosome.

H_1: The epigenetic marking context is not homogeneous among histones chromosome by chromosome.

From the test, we concluded that the epigenetic mark context is not homogeneous in any chromosome, thus rejecting the H_0.

As to the Cramer's V values, these are similar for all chromosomes, except for chromosome Y which has a higher heterogeneity effect. The respective analysis of residues, was conducted for all chromosomes, for nucleotides, dinucleotides, trinucleotides and tetranucleotides, but here we show only a few chromosomes, as an example, such as chromosome 1 for nucleotides (see in figure 19), that illustrates well what happened in general the remaining chromosomes. Again in this analysis, the chromosome Y is the exception (see in figure 22).

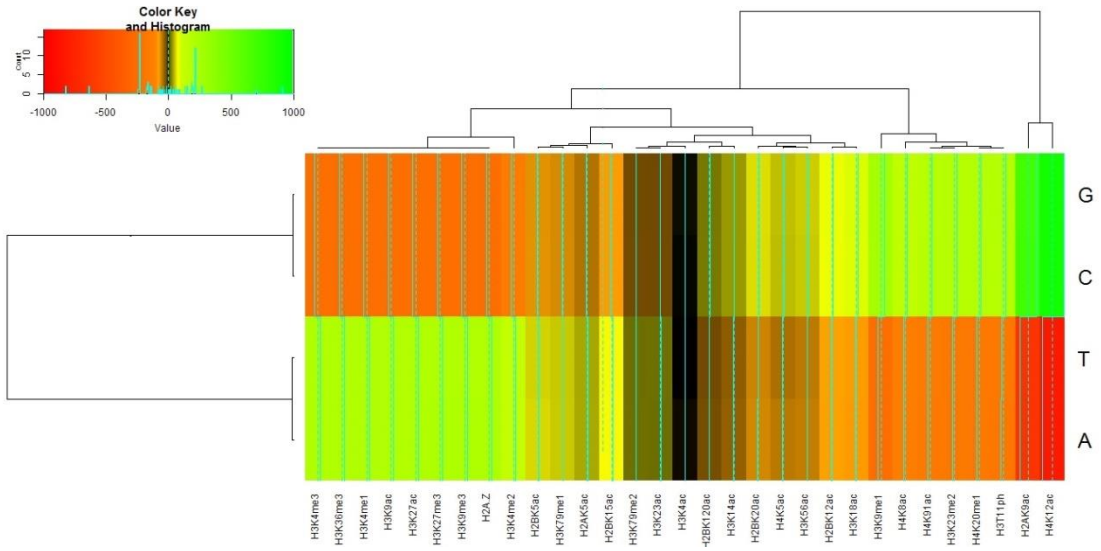


Figure 19: Heatmap of global union for nucleotide context of chromosome 1. The chromosomes have an identical profile, highlighting H4K12ac H2AK9ac that has a preference for contexts enriched in G and C. The remaining histones have a more or less homogenous behavior.

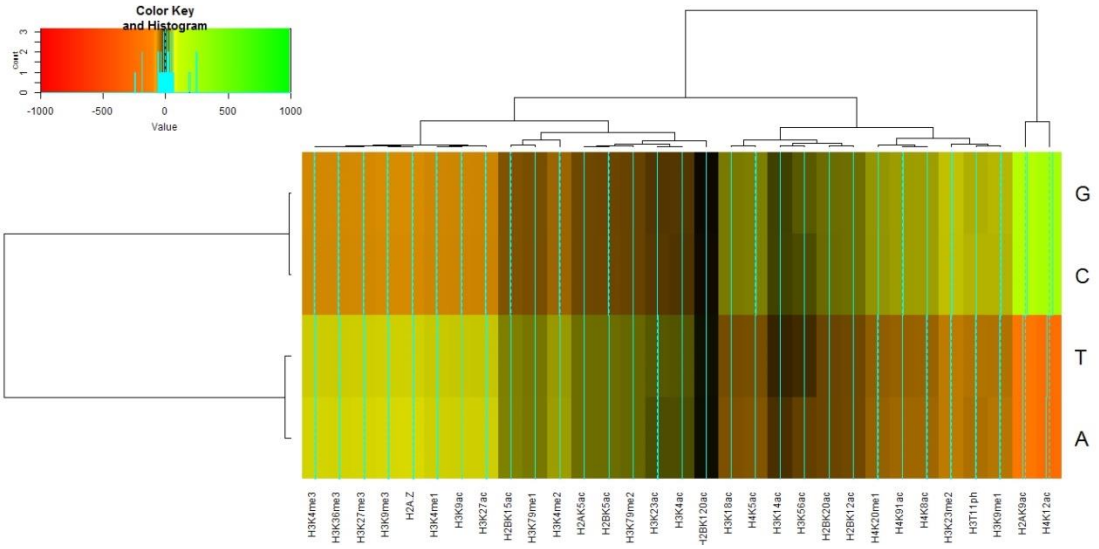


Figure 20: Heatmap of global union for nucleotide contexts of chromosome 22. We can see again that the histones H4K12ac H2AK9ac that have a preference for contexts enriched in G and C. The remaining contents histones have a more or less homogeneous behavior.

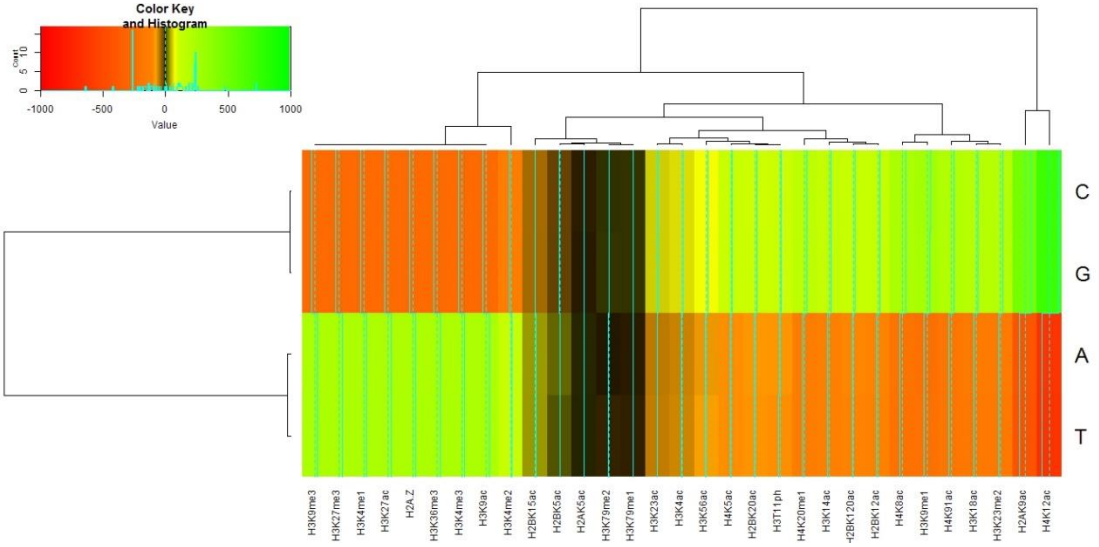


Figure 21: Heatmap of global union for nucleotide contexts of chromosome X. It shows that the histones H4K12ac H2AK9ac that have a preference for contexts enriched in G and C. The remaining contents histones have a more or less homogeneous behavior.

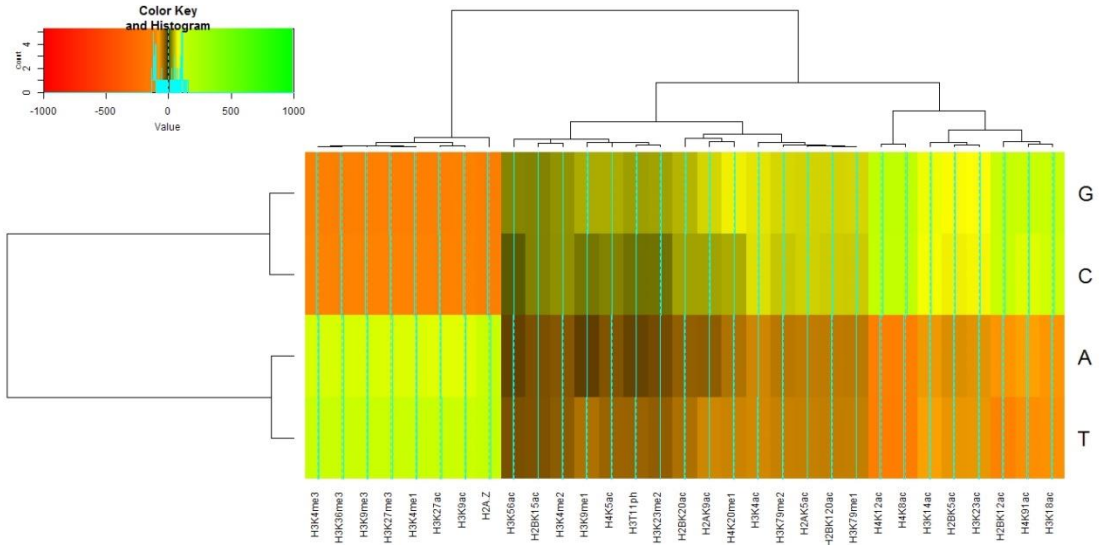


Figure 22: Heatmap of global union for nucleotide contexts of chromosome Y. Observed three groups where the content of the histones is homogeneous, one of them have preference for contexts enriched in A and T; another group have preference for contexts enriched in G and C; The third group have a homogeneous behavior.

In general, all chromosomes have an identical profile (e.g. figures 19, 20 and 21), which highlights H4K12ac and H2AK9ac that have a preference for nucleotides G and C. It is possible to identify three relevant clusters, one formed by H4K12ac and H2AK9ac; another cluster that although having more homogeneous preference, show an overall preference for G and C; and third cluster that is also more homogeneous, but contains certain histone modifications with preference for A and T.

As stated previously, there is one exception to this profile that corresponds to the chromosome Y, figure 22. Here we observed three clusters where the content of the modifications is homogeneous, one is formed by H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K4me1, H3K9me3, H3K27ac, H3K9ac and H2AZ, and have preference for A and T; another is formed by H4K8ac, H4K12ac, H3K14ac, H2BK5ac, H3K23ac, H2Bk12ac, H4K91ac and H3K18 ac, and shows are preference for nucleotides G and C; The third cluster is formed by the remaining histones, and has a homogeneous behavior.

3.4.5. Analysis of different histone modifications

It was also our intention to understand in what epigenomes, the chromosomes are more different from each other. To this end, we apply the chi-square test and the measure of association Cramer's V for the histone union and histone intersection, see in annex 6.

This analysis was done with the objective of responding to this test hypotheses:

H_0: The epigenetic marking context is homogeneous among chromosomes for each histone modification.

H_1: The epigenetic marking context is not homogeneous among chromosomes for each histone modification.

The results of this test and its values show in annex 6, reveals that the occurrence epigenetic marking is significantly heterogeneous among chromosomes. Regarding the values of Cramer's V, we can see that there is a greater association at the histone intersection level than the histone union of histones.

To better understand the results for the histone union, the heatmap for nucleotides, dinucleotide, trinucleotide and tetranucleotides for all histone modifications was built. However, we show only one example to illustrate the common trend of in this analysis (see in figure 23). In relation to histone intersection, we show several examples, as not all histones had the same profile (see in figures 24, 25, 26 and 27).

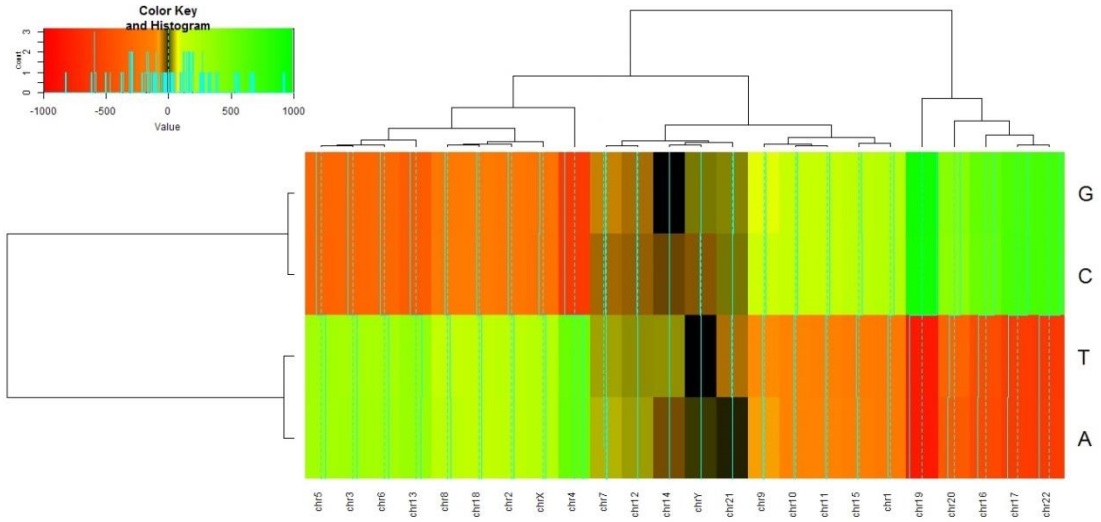


Figure 23: Heatmap of H2AK5ac nucleotide preference, using the histone union dataset. There are five groups in which the content of chromosomes is homogeneous: two groups, although separated, have preference for nucleotides G and C; two groups, also separated, have preference for nucleotides A and T; and one more group has a homogenous behavior.

The heatmap in figure 23 features three clusters in which the content of chromosomes is homogeneous: one of them is formed by chromosomes 9, 10, 11, 15, 1, 19, 20, 16, 17 and 22, and have preference for nucleotides G and C; another one is formed by chromosomes 5, 3, 6, 13, 8, 18, 2, X and 4 which prefer nucleotides A and T;

and, finally, another cluster that homogenous behavior. Also chromosomes 19 and 4 exhibited a globally different behavior in comparison to others.

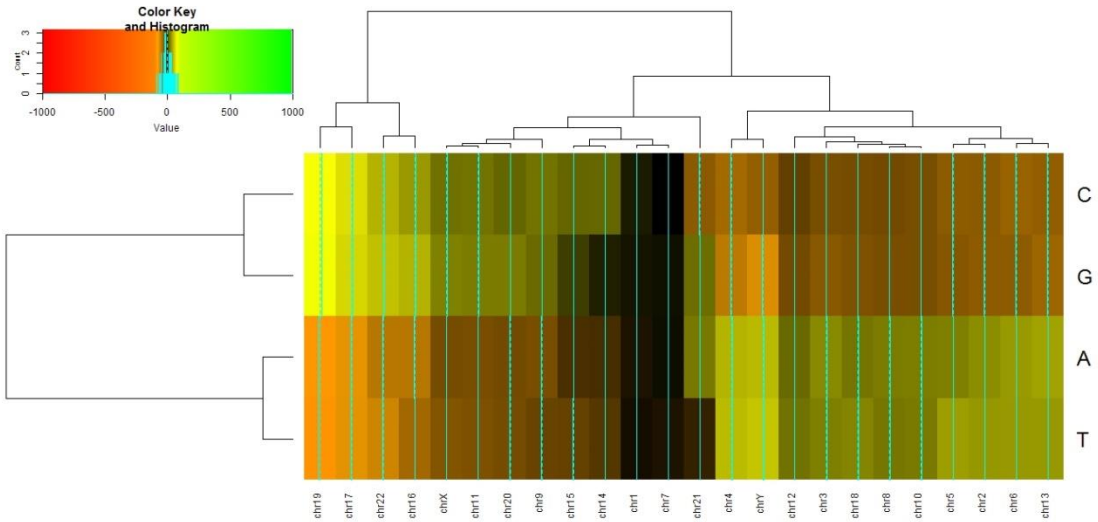


Figure 24: Heatmap of H2AZ nucleotide preference, using the histone intersection datasets. There are three clusters: one that prefers nucleotides G and C; another that prefers nucleotides A and T; and a last one that shows a homogeneous behavior.

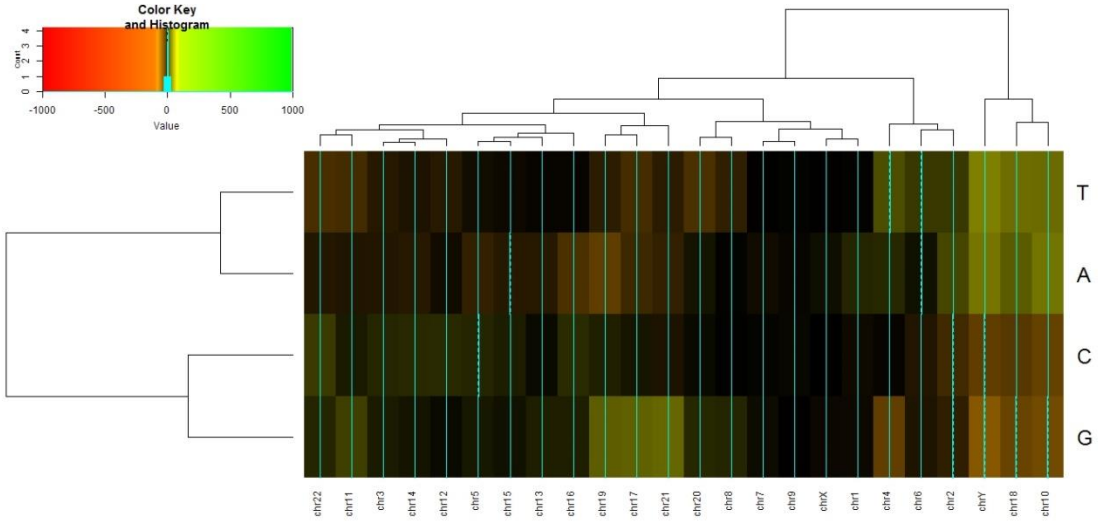


Figure 25: Heatmap of H2BK5ac nucleotide preference, using the histone intersection datasets. There are two cluster homogeneous, one of which has preferably at the content of the nucleotides A and T; and another which has a preference for the homogeneous content.

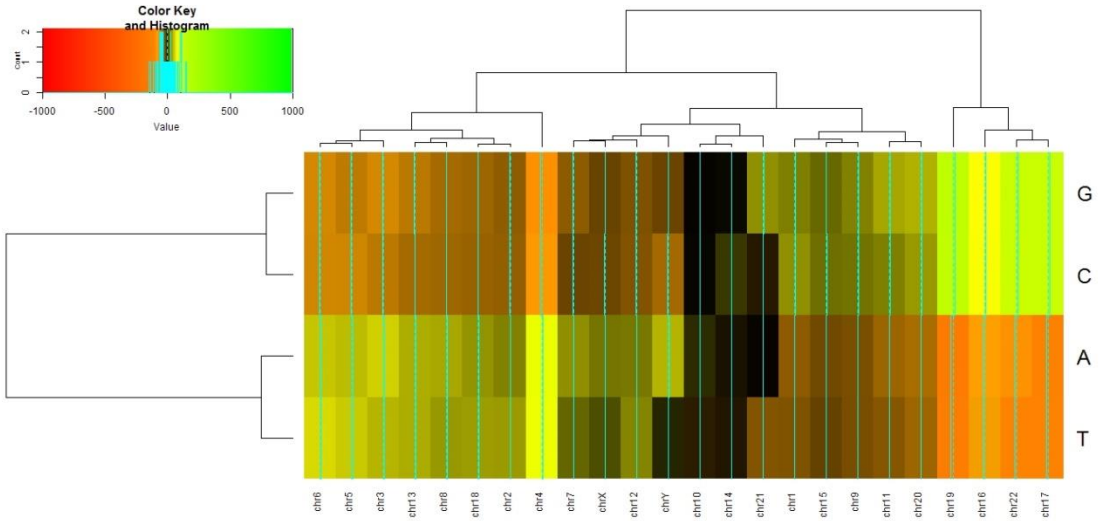


Figure 26: Heatmap of H2BK20ac nucleotide preference, using the histone intersection datasets. There are two clusters: one of which prefers nucleotides G and C; other that prefers nucleotides A and T.

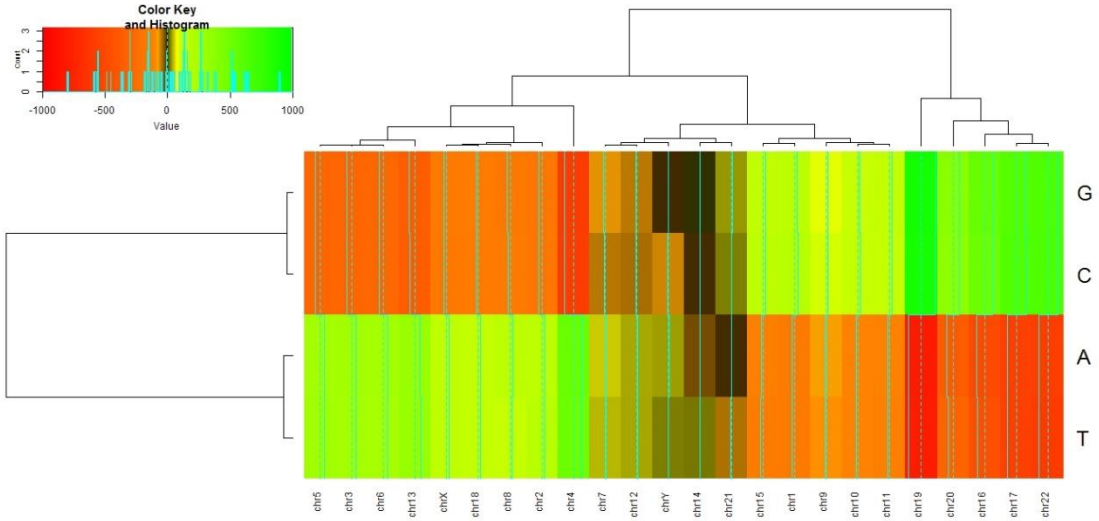


Figure 27: Heatmap of H4K5ac nucleotide preference, using the histone intersection datasets. We can see three clusters: one of which prefers nucleotides G and C; another that prefers nucleotides A and T; and a last one that shows a homogeneous behavior.

The epigenetic marking context is not homogeneous among chromosomes for each histone modification, as already seen in annex 6, and these heatmaps can confirm those values.

In heatmap of figure 24 relative to H2AZ, shows three clusters that present homogeneity: one formed by chromosomes 19, 17, 22 and 16 having preference for nucleotides G and C; another formed by chromosomes 4, Y, 12, 3, 18, 8, 10, 5, 2, 6 and 13 having preference for nucleotides A and T; and a last cluster formed by the remaining chromosomes that did not show a marked preference by any nucleotide.

Figure 25 shows the heatmap of H2BK5ac, where only two clusters are homogeneous, one of which shows preference for nucleotides A and T, formed by chromosomes Y, 8 and 10; and another one formed by the remaining chromosomes, with no marked preference.

As to H2BK20ac, figure 26, two clusters were formed: one formed by chromosomes 17, 16, 22 and 19, with preference for nucleotides G and C; another one formed by the remaining chromosomes, with preference for nucleotides A and T.

One last example that proves that the marking epigenetic context is not homogeneous among chromosomes when each histone modification is considered is the case of H4K5ac (see in figure 27). We can identify three clusters: one of which is formed by chromosomes 19, 20, 16, 17, 22, 11, 10, 9, 1 and 15 which prefers nucleotides G and C; another cluster formed by the chromosomes 5, 3, 13, X, 6, 18, 8, 2 and 4 that has preference for nucleotides A and T; and finally a cluster formed by the remaining chromosomes with no major a preference for any nucleotide context.

Another analysis was performed in order to verify of the epigenetic marking context for each histone modification was equal to the unmarked control sequences. To this end, again we applied the chi-square test and the Cramer's V for nucleotides, dinucleotides, trinucleotides and tetranucleotides (see annex 7).

The hypotheses tested this time were:

H_0: The epigenetic marking context for each histone is equal to the epigenetic marking context for control sequences.

H_1: The epigenetic marking context for each histone is not equal to the epigenetic marking context for control sequences.

Through this test we were able to conclude in favour of H1, which suggest that specific contexts somehow originate specific histone modification.

Despite the values of Cramer’s V being low, we still performed the analysis of residues. However, most of the results can be considered negligible. We will just show some corresponding to the strongest Cramer's V (see in figures 28, 29 and 30).

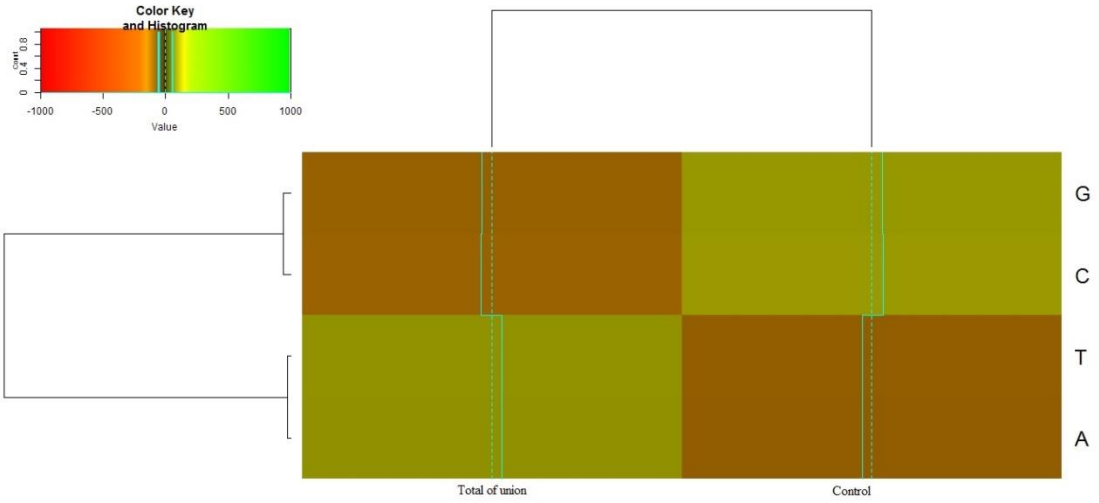


Figure 28: Heatmap for H2AZ at the level of nucleotides. This modification has preference for Ts.

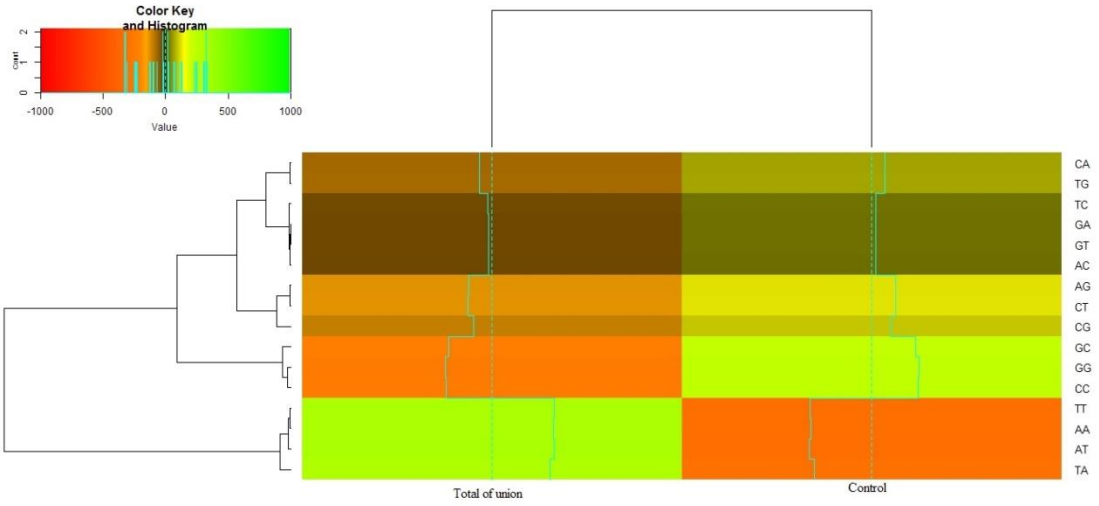


Figure 29: Heatmap for H2AK9ac at the level of dinucleotides. There is a preference for the TT, AA, AT and TA contexts.

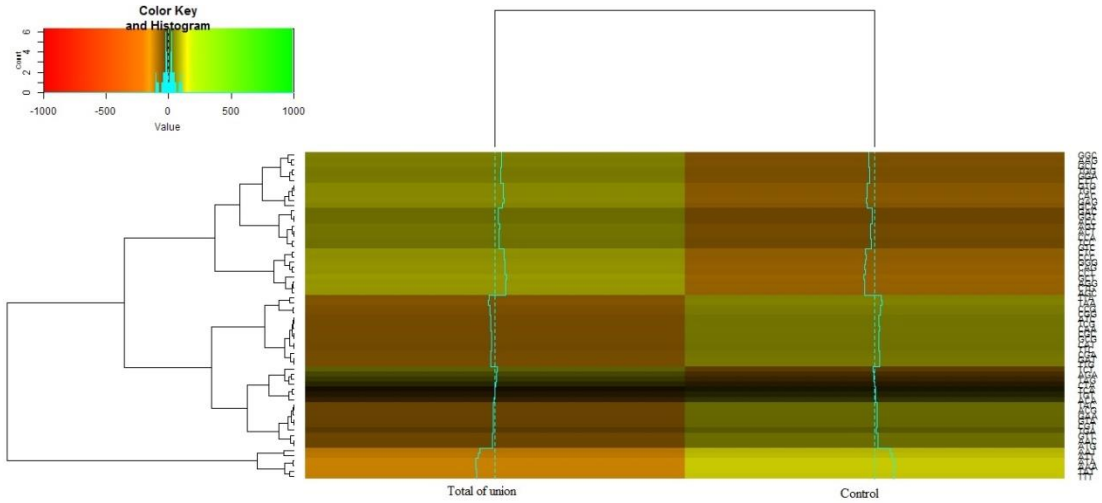


Figure 30: Heatmap for H3T11ph at the level of trinucleotides. There is a preference for TAA contents.

As we can see in the heatmap of figure 28, for H2AZ, there was a preference for T and A, while the control had preference for C and G. Meaning, that fragments enriched in T and A, had a higher tending to have.

In relation to H2AK9ac (see in figure 29) there's a preference for TT, AA, AT and TA, while the control had preference for GC, GG and CC.

As to H3T11ph, figure 30, we see a preference for the trinucleotide GCT. In the control the preference is for TTT, TAT, AAA, ATA and AAT.

What can we conclude from the above heatmaps is that each modifications its own preference. Therefore, it becomes possible to identify specific genomic contexts in histone modifications (table 12):

Histone modifications	Identify specific genomic contexts
H2AZ	TTA and TAA
H2AK5ac	
H2BK5ac	
H2Bk15ac	
H3K4ac	
H3K4me1	
H3K4me2	
K3K4me3	
H3K9ac	

H3K9me3 H3K23ac H3K27ac H3K27me3 H3K36me3 H3K79me1 H3K79me2	
H2AK9ac	GGG, GCC and GGC
H2BK12ac H2BK20ac H3K14ac H3K18ac H3K56ac H4K12ac	GCT and CTC
H2BK120ac	TTA, TAA, CTC and GCT
H3K9me1	AGG, GCT, CTG and CCT
H3K23me2	AGG, CTC and GCT
H3T11ph	GCT
H4K5ac	GCT, CTC and TTA
H4K12ac	GGG, GGC, CCC, GCC and GAG
H4K20me1	CAG, CAC, AGG, CTC, GCT and CCT
H4K91ac	GCT, CTT and AGG

Table 12: Identify specific genomic contexts in histone modifications.

4. Discussion

The type of work presented here is a pilot study. That is, a test small, of procedures, materials and methods proposed for a given search.

Throughout this dissertation we presented results obtained from the application of various statistical methodologies to the study of the human genome. The main objective was to study the epigenetic signals of the human genome, which can identify motifs for each histone modification. Thus, understanding the relationship between DNA sequences and the occurrence of epigenetic marking.

In this study the following methods were used:

- Contingency tables analysis;
- Measure of association Cramer's V;
- Heatmaps;
- Dendograms.

Through contingency tables, we used statistical chi-square test, to reject the hypothesis of independence between the DNA content and the occurrence of epigenetic marking and we obtained small amounts of association. In conclusion, there is dependence between the content and the occurrence of epigenetic marking.

Through heatmaps, we could identify specific genomic contexts for each histone modification. One of the strongest contexts is that TTA and TAA trinucleotides are present mainly in regions of H2 and H3 histone modification, for both acetylation and methylation. However, there are other histone modifications that have other enriched motifs, as shown in the results. So, with this analysis, it was possible to predict the occurrence of a modification from the nucleotide of the region context.

The histone modifications that showed highest differences between chromosomes histones were H4K12ac and H2AK9ac. H4K12ac is related to memory in its consolidation (annex 4) since there is evidence that the epigenome regulates the consolidation of information processed in long-term memory. In relation to H2AK9ac, which is also changed in most chromosomes, in addition to other histone modifications changed, , such as H4K8ac, in the chromosome Y (which is the least homogeneous chromosome), they are all connected to the activation / inactivation of the transcription (annex 4). As mentioned earlier, epigenetic mechanisms, such as DNA methylation,

histone modifications and ncRNAs, pattern the activation / inactivation of genes being powerful regulators of gene activity, RNA transcription and protein homeostasis. Therefore, one would expect that certain histone were changed more than others.

Through the dendrogram analysis, it was possible to create groups by the type of histones (H2, H3 and H4) and the type of modification (acetylation or methylation) based on the samples behavior. Interestingly, this separates groups including transcription-activating histone modifications (normally acetylations) and transcription-inactivating owes (normally methylation).

We only analyze nucleotide, dinucleotide, trinucleotide and tetranucleotide contexts, the latter being the most informative context to take conclusions because it is the word size of nucleotides is greater. From the results, we can say that increasing the word size of contexts, more information and conclusions could be drawn. The word size of nucleotides was a limitation of this analysis.

4.1. Future work

According to the results, it will be interesting to do the following future work:

- In terms of genomic location, to evaluate interactions among epigenomes and motifs position in the fragment;
- To performed this analysis increasing the word size of contexts, and thus relating with to specific genomic contexts in histone modification and in chromosomes;
- To study the effect of variables present in the metadata, such as age, sex, anatomy, epigenome class, ethnicity and solid/liquid status;
- The results now obtained with this information about the expression of genes, to select the relevant set points and study them separately.

5. Bibliography

1. Azevedo C. Biologia Celular e Molecular. 4ª Edição. Lidel; 2005. 141-191 p.

2. No Title [Internet]. Available from: http://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/rna2.html.

3. Ng RK, Gurdon JB. Epigenetic inheritance of cell differentiation status. Cell Cycle. 2008;7(9):1173–7.

4. Roloff TC, Nuber UA. Chromatin, epigenetics and stem cells. Eur J Cell Biol. 2005;84(2-3):123–35.

5. Probst A V., Dunleavy E, Almouzni G. Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol [Internet]. 2009;10(3):192–206. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19234478

6. WHO. Genetics, genomics and the patenting of DNA: review of potential implications for health in developing countries. World Heal Organ [Internet]. 2005; Available from: http://www.who.int/genomics/FullReport.pdf?ua=1

7. Bernstein BE, Meissner A, Lander ES. The Mammalian Epigenome. Cell. 2007;128(4):669–81.

8. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. Analysis of Dynamic Changes in Posttranslational Modifications of Human Histones during Cell Cycle by Mass Spectrometry. NIH Public Access. 2013;28(10):1045–8.

9. Scholz B, Marschalek R. Epigenetics and blood disorders. Br J Haematol. 2012;158(3):307–22.

10. Chen T, En L. Structure and Function of Eukaryotic DNA Methyltransferases. Curr Top Dev Biol. 2004;60:55–9.

11. No Title [Internet]. Available from: http://epidemiologiamolecular.com/nucleo-y-nucleolo/

12. No Title [Internet]. Available from: http://bioilogicos.blogspot.pt/2012/03/dna-cromatina-cromossomo-gene.html

13. No Title [Internet]. Available from: https://probiokelinton.wordpress.com/2010/04/28/o-nucleo-da-celula-parte-2/

14. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.

15. Macdonald N, Welburn JPI, Noble MEM, Nguyen A, Yaffe MB, Clynes D, et al. Molecular basis for the recognition of phosphorylated and phosphoacetylated histone H3 by 14-3-3. Mol Cell. 2005;20(2):199–211.

16. Zaratiegui M, Irvine D V., Martienssen RA. Noncoding RNAs and Gene Silencing. Cell. 2007;128(4):763–76.

17. Kouzarides T. Chromatin Modifications and Their Function. Cell. 2007;128(4):693–705.

18. Jin B, Keith D. Robertson. DNA Methyltransferases (DNMTs), DNA Damage Repair, and Cancer. NIH Public Access. 2013;1–25.

19. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247–57.

20. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4:143–53.

21. Ghildiyal, M; Zamore PD. Small silencing RNAs: an expanding universe. Mol Pharmacol. 2009;10(2):94–108.

22. Carthew, Richard W. and Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009;136(4):642–55.

23. Cerutti H, Casas-Mallano JA. Histone H3 phosphorylation: Universal code or lineage specific dialects? Epigenetics. 2009;4(2):71–5.

24. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. 2014;132(4):567–82.

25. Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358:1148–59.

26. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A [Internet]. 1992;89(5):1827–31. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=48546&tool=pmcentrez&rendertype=abstract

27. Tollefsbol TO. Epigenetics Protocols. Platelets. 2004;287(18):316.

28. Callinan PA, Feinberg AP. The emerging science of epigenomics. Hum Mol Genet. 2006;15 Spec No(1):95–101.

29. Bird AP. Use of restriction enzymes to study eukaryotic DNA methylation. II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. J Mol Biol. 1978;118(1):49–60.

30. Clark SJ, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. Nucleic Acids Res. 1994;22(15):2990–7.

31. Hajkova P, el-Maarri O, Engemann S, Oswald J, Olek A, Walter J. DNA-methylation analysis by the bisulfite-assisted genomic sequencing method.

Methods Mol Biol [Internet]. 2002;200(4):143–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11951649

32. Olek A, Oswald J, Walter J. A modified and improved method for bisulphite based cytosine methylation analysis. Nucleic Acids Res. 1996;24(24):5064–6.

33. Paulin R, Grigg GW, Davey MW, Piper AA. Urea improves efficiency of bisulphite-mediated sequencing of 5 0 methylcytosine in genomic DNA. Nucleic Acids Res. 1998;26(21):5009–10.

34. Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. NIH Public Access. 2015;18(11):1492–501.

35. Pelizzola M, Koga Y, Urban AE, Dindot S V, Person R, Strivens M, et al. MEDME : An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MEDME : An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-e. 2009;1652–9.

36. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. 2000;9(16):2395–402.

37. Ozanne SE, Constância M. Mechanisms of disease: the developmental origins of disease and the role of the epigenotype. Nat Clin Pract Endocrinol Metab [Internet]. 2007;3(7):539–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17581623

38. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature [Internet]. 2015;518(7539):317–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25693563

39. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2013;28(10):1045–8.

40. Varuzza L. Introdução à análise de dados de sequenciadores de nova geração. Leonardo Varuzza's Site [Internet]. 2013;1–76. Available from: http://lvaruzza.com/files/apostila_bioinfo_2.0.1.pdf

41. No Title [Internet]. Available from: http://www.roadmapepigenomics.org/data/tables/all

42. Venables WN, Smith DM. An Introduction to R. 2016;4.

43. Aboyoun P, Gentleman R, Debroy S, Rmpi E. Package " Biostrings ." 2016;

44. Schwartz M. Package " WriteXLS ." 2015;2003:1–8.

45. Dragulescu MAA. Package " xlsx ." 2015;

46. Carlson M. An Introduction to the GenomicRanges Package GRanges : Genomic Ranges. 2016;1–22.

47. Mcgough JJ, Faraone S V. ESTIMATING THE SIZE OF TREATMENT EFFECTS : Moving Beyond P Values. 2009;6(10):21–9.

48. Mchugh ML. Lessons in biostatistics The Chi-square test of independence. 2013;23(2):143–9.

49. Ugoni A, Walker BF. THE CHI SQUARE TEST An introduction. 1995;4(3):61–4.

50. Statistical N, Ncss S, Reserved AR. Hierarchical Clustering / Dendrograms. :1–15.

51. Vaquero A, Loyola A, Reinberg D. The constantly changing face of chromatin. Sci Aging Knowl Env. 2003;14.

52. Martin C, Zhang Y. The diverse functions of histone lysine methylation. Nat Rev. 2005;6(11):838–49.

53. Zhang Y, Reinberg D. Transcription regulation by histone methylation: Interplay between different covalent modifications of the core histone tails. Genes Dev. 2001;15(18):2343–60.

54. Hochstrasser M. Ubiquitin-Dependent Protein. 1996;(93). Available from: http://www.annualreviews.org/doi/pdf/10.1146/annurev.genet.30.1.405

55. Hershko A, Ciechanover A. The ubiquitin system. Annu Rev Biochem. 1998;67:425–79.

56. Olson MO, Prestayko AW, Olson J, Prestayko AW, Busch H. Isolation and Characterization of Protein Chromosomal Protein *. J Biol Chem. 1975;(18):7182–7.

57. Osley MA. Regulation of histone H2A and H2B ubiquitylation. Briefings Funct Genomics Proteomics. 2006;5(3):179–89.

58. Weake VM, Workman JL. Histone Ubiquitination: Triggering Gene Activity. Mol Cell. 2008;29(6):653–63.

59. Osley MA. H2B ubiquitylation: the end is in sight. Biochim Biophys Acta. 2004;1677(1-3):74–8.

60. Henry KW, Berger SL. Trans-tail histone modifications: wedge or bridge? Nat Struct Biol. 2002;9(8):565–6.

61. Cao J, Yan Q. Histone ubiquitination and deubiquitination in transcription, DNA damage response, and cancer. Front Oncol [Internet]. 2012;2(March):26. Available from: /pmc/articles/PMC3355875/?report=abstract

62. Turjanski AG, Vaqué JP, Gutkind JS. MAP kinases and the control of nuclear events. Oncogene [Internet]. 2007;26(22):3240–53. Available from:

http://www.nature.com/doifinder/10.1038/sj.onc.1210415\nhttp://www.ncbi.nlm.nih.gov/pubmed/17496919

63. Johansen KM, Johansen J. Regulation of chromatin structure by histone H3S10 phosphorylation. Chromosom Res. 2006;14(4):393–404.

64. Tomari Y, Zamore PD. Perspective: Machines for RNAi. Genes Dev. 2005;19(5):517–29.

65. Negrini M, Nicoloso MS, Calin GA. MicroRNAs and cancer-new paradigms in molecular oncology. Curr Opin Cell Biol. 2009;21(3):470–9.

66. Ishimoto H, Jaffe RB. Development and Function of the Human Fetal Adrenal Cortex: A Key Component in the Feto-Placental Unit. Endocr Rev. 2011;32:317–55.

67. Rainey WE, Rehman KS, Carr BR. The Human Fetal Adrenal: Making Adrenal Androgens for Placental Estrogens. Semin Reprod Med. 2004;22:327–36.

68. Guyton AC, Hall JE. Tratado de Fisiologia Médica. 9ºEdição ed. Guanabara Koogon; 1997.

69. Drake RL, Vogl W, Mitchelll AWM. Gray-Anatomia para Estudantes. Elsevier; 2005.

70. Farrar W. Cancer Stem Cells. N Engl J Med. 2009;355(12):1–191.

71. Kaur S, Singh G, Kaur K. Cancer stem cells: An insight and future perspective. J Can Res Ther. 2014;10:846–52.

72. Arosa FA, Cardoso EM, Pacheco FC. Fundamentos de Imunologia. 2º Edição. Lidel, editor. 2012.

73. Leeper NJ, Hunter AL, Cooke JP, Nicholas J. Leeper, Arwen L.Hunter and JPC. Stem cell therapy for vascular regeneration: adult, embryonic, and induced pluripotent stem cells. Circulation. 2010;122(5):517–26.

74. Comyn O, Lee E, Maclaren RE. Induced pluripotent stem cell therapies for retinal disease. Curr Opin Neurol. 2010;23(1):4–9.

75. Netter FH. Atlas de Anatomia Humana. 5ºEdição ed. Elsevier, editor. 2011.

76. Junqueira LC, Carneiro J. Histologia Básica. 11º Edição. Guanabara Koogon; 2008.

77. Fausto CSCDV, Chammas MC, Saito ODC, Garcia MRT, Juliano AG, Simões CA, et al. Timo: caracterização ultra-sonográfica. Radiol Bras. 2004;37(3):207–10.

78. Hake SB. Histone H2A variants in nucleosomes and chromatin : more or less stable ? 2012;40(21):10719–41.

79. Kalocsay M, Hiller NJ, Jentsch S. Article SUMO-H2A . Z-Dependent Chromosome Fixation in Response to a Persistent DNA Double-Strand Break. Mol Cell

[Internet]. Elsevier Ltd; 2009;33(3):335–43. Available from: http://dx.doi.org/10.1016/j.molcel.2009.01.016

80. Sarcinella E, Zuzarte PC, Lau PNI, Draker R, Cheung P. Monoubiquitylation of H2A . Z Distinguishes Its Association with Euchromatin or Facultative Heterochromatin . 2007;27(18):6457–68.

81. Thambirajah AA, Dryhurst D, Ishibashi T, Li A, Maffey AH, Ausio J. H2A . Z Stabilizes Chromatin in a Way That Is Dependent on Core Histone Acetylation *. 2006;281(29):20036–44.

82. Zlatanova J, Thakar A. Review H2A . Z : View from the Top. 2008;(February):166–79.

83. No C. Histone H2AK5ac antibody (pAb). 32(0):3–4.

84. Wang Z, Zang C, Rosenfeld J a, Schones DE, Cuddapah S, Cui K, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet. 2009;40(7):897–903.

85. Singh P, Cho J, Tsai SY, Rivas GE, Larson GP, Szabó PE. Coordinated allele-specific histone acetylation at the differentially methylated regions of imprinted genes. Nucleic Acids Res. 2010;38(22):7974–90.

86. Vlaicu SI, Tegla C a., Cudrici CD, Fosbrink M, Nguyen V, Azimzadeh P, et al. Epigenetic modifications induced by RGC-32 in colon cancer. Exp Mol Pathol [Internet]. 2010;88(1):67–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19883641

87. Golebiowski F, Kasprzak K. Inhibition of core histones acetylation by carcinogenic nickel(II). Mol Cell Biochem. 2005;279:1333–9.

88. Shlyueva D, Stampfel G, Stark A. Transcriptional enchancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15:272–86.

89. Liang G, Lin JCY, Wei V, Yoo C, Cheng JC, Nguyen CT, et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A. 2004;101:7357–62.

90. Barlési F, Giaccone G, Gallegos-Ruiz MI, Loundou A, Span SW, Lefesvre P, et al. Global histone modifications predict prognosis of resected non-small-cell lung cancer. J Clin Oncol. 2007;25(28):4358–64.

91. Seligson DB, Horvath S, McBrian M a, Mah V, Yu H, Tze S, et al. Global levels of histone modifications predict prognosis in different cancers. Am J Pathol. 2009;174(5):1619–28.

92. Coffee B, Zhang F, Ceman S, Warren ST, Reines D. Histone modifications depict an aberrantly heterochromatinized FMR1 gene in fragile x syndrome. Am J Hum Genet. 2002;71(4):923–32.

93. Mcgarvey KM, Neste L Van, Cope L, Ohm JE, James G, Criekinge W Van, et al. Hypermethylated Genes in Colon Cancer Cells. Cancer. 2009;68(14):5753–9.

94. Flanagan JF, Mi LZ, Chruszcz M, Cymborowski M, Clines KL, Kim Y, et al. Double chromodomains cooperate to recognize the methylated histone H3 tail. Nature. 2005;438:1181–5.

95. Wang H, Ramakrishnan A, Fletcher S, Prochownik E V, Genetics M. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature. 2015;2(2):91–5.

96. Nishioka K, Chuikov S, Sarma K, Erdjument-Bromage H, Allis CD, Tempst P, et al. Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. Genes Dev. 2002;16(4):479–89.

97. Schneider R, Bannister AJ, Weise C, Kouzarides T. Direct binding of INHAT to H3 tails disrupted by modifications. J Biol Chem. 2004;279(23):23859–62.

98. Lauberth SM, Nakayama T, Wu X, Ferris AL, Tang Z, Hughes SH, et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell [Internet]. Elsevier Inc.; 2013;152(5):1021–36. Available from: http://dx.doi.org/10.1016/j.cell.2013.01.052

99. Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC Genomics [Internet]. 2012;13:424. Available from: BMC Genomics

100. Metzger E, Wissmann M, Yin N, Müller J, Schneider R, Peters A, et al. LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. Nature. 2005;437:436–9.

101. Sims J, Houston S, Magazinnik T, Rice J. A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. J Biol Chem. 2006;

102. Kim J, Kim H. Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. ILAR J [Internet]. 2012;53(3-4):232–9. Available from: http://ilarjournal.oxfordjournals.org/content/53/3-4/232.long

103. Agalioti T, Chen G, Thanos D. Deciphering the transcriptional histone acetylation code for a human gene. Cell. 2002;

104. Tzao C, Tung H, Jin J, Sun G, Hsu H, Chen B, et al. Prognostic significance of global histone modifications in resected squamous cell carcinoma of the esophagus. Mod Pathol. 2009;22:252–60.

105. Liu BL, Cheng JX, Zhang X, Wang R, Zhang W, Lin H, et al. Global histone modification patterns as prognostic markers to classify glioma patients. Cancer

Epidemiol Biomarkers Prev. 2010;19(11):2888–96.

106. Tsai W, Wang Z, Yiu TT, Akdemir KC, Xia W, Winter S, et al. TRIM24 links a noncanonical histone signature to breast cancer Wen-Wei. Nature. 2011;468(7326):927–32.

107. Vandamme J, Sidoli S, Mariani L, Friis C, Christensen J, Helin K, et al. H3K23me2 is a new heterochromatic mark in Caenorhabditis elegans. Nucleic Acids Res. 2015;

108. Tie F, Banerjee R, Stratton CA, Prasad-Sinha J, Stepanik V, Zlobin A, et al. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. Development. 2009;136:3131–41.

109. Kuzmichev A, Nishioka K, Erdjument-bromage H, Tempst P, Reinberg D. multiprotein complex containing the Enhancer of Zeste protein Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. 2002;2893–905.

110. Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K. et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006;441:349–53.

111. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. 2006;1123–36.

112. Rougeulle C, Chaumeil J, Sarma K, Allis CD, Reinberg D, Avner P, et al. Differential Histone H3 Lys-9 and Lys-27 Methylation Profiles on the X Chromosome †. 2004;24(12):5475–84.

113. Magill JC, Byl MF, Goldwaser B, Instructor MP, Yates B, Morency JR, et al. IDENTIFICATION OF HISTONE H3 LYSINE 36 ACETYLATION AS A HIGHLY CONSERVED HISTONE MODIFICATION*. J Biol Chem. 2010;3(1):1–19.

114. Fnu S, Williamson EA, De Haro LP, Brenneman M, Wray J, Shaheen M, et al. Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. Proc Natl Acad Sci [Internet]. 2011;108(2):540–5. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21187428&retmode=ref&cmd=prlinks\npapers2://publication/doi/10.1073/pnas.1013571108

115. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira- OM, Misteli T. Regulation of Alternative Splicing by Histone Modifications. Science (80-). 2010;327(5968):996–1000.

116. Das C, Lucia MS, Hansen KC, Tyler JK. CBP / p300-mediated acetylation of histone H3 on lysine 56. Nature. 2009;459(7243):113–7.

117. Wang H, Ramakrishnan A, Fletcher S, Prochownik E V, Genetics M. Cell cycle-dependent deacetylation of telomeric histone H3 lysine K56 by human SIRT6. Cell Cycle. 2015;2(2):2664–6.

118. Vempati RK, Jayani RS, Notani D, Sengupta A, Galande S, Haldar D. p300-mediated Acetylation of Histone H3 Lysine 56 Functions in DNA Damage Response in Mammals. J Biol Chem. 2010;28553–64.

119. Rossodivita AA, Boudoures AL, Mecoli JP, Steenkiste EM, Karl AL, Vines EM, et al. Histone H3 K79 methylation states play distinct roles in UV-induced sister chromatid exchange and cell cycle checkpoint arrest in Saccharomyces cerevisiae. Nucleic Acids Res. 2014;42(10):6286–99.

120. Fu H, Maunakea AK, Martin MM, Huang L, Zhang Y, Ryan M, et al. Methylation of Histone H3 on Lysine 79 Associates with a Group of Replication Origins and Helps Limit DNA Replication Once per Cell Cycle. PLoS Genet. 2013;9(6):1–14.

121. Feng Q, Wang H, Ng H, Erdjument-Bromage H, Tempst P, Struhl K, et al. Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. Curr Biol. 2002;

122. Preuss U, Landsberg G, Scheidtmann KH. Novel mitosis-speci ® c phosphorylation of histone H3 at Thr11 mediated by Dlk / ZIP kinase. 2003;31(3):878–85.

123. Metzger E, Yin N, Wissmann M, Kunowska N, Fischer K, Friedrichs N, et al. Phosphorylation of histone H3 at threonine 11 establishes a novel chromatin mark for transcriptional regulation. Nat Cell Biol. 2010;10(1):53–60.

124. Thatcher TH, Gorovsky MA. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res [Internet]. 1994;22(2):174–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=307768&tool=pmcentrez&rendertype=abstract

125. Zhao R, Nakamura T, Fu Y, Lazar Z, Spector DL, Harbor CS, et al. Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation. Nat Cell Biol. 2012;13(11):1295–304.

126. Park CS, Rehrauer H, Mansuy IM. Genome-wide analysis of H4K5 acetylation associated with fear memory in mice. BMC Genomics. 2013;14.

127. Peleg S, Sananbenesi F, Zovoilis A, Burkhardt S, Bahari-Javan S, Agis-Balboa RC, et al. Altered histone acetylation is associated with age-dependent memory impairment in mice. Science (80-). 2010;328:753–6.

128. Larschan E, Winston F. The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4. Genes Dev. 2001;15(15):1946–56.

129. van der Heijden GW, Derijck AA, Ramos L, Giele M, van der Vlag J, Boer P.

Transmission of modified nucleosomes from the mouse male germline to the zygote and subsequent remodeling of paternal chromatin. Dev Biol. 2006;298:458–69.

130. Paradowska AS, Miller D, Spiess A., Vieweg M, Cerna M, Bartkuhn M, et al. Genome wide identification of promoter binding sites for H4K12ac in human sperm and its relevance for early embryonic development. Epigenetics. 2012;7:1057–70.

131. Beck DB, Burton A, Oda H, Ziegler-Birling C, Torres-Padilla ME, Reinberg D. The role of PR-Set7 in replication licensing depends on Suv4-20h. Genes Dev. 2012;26(23):2580–9.

132. Kuo AJ, Song J, Cheung P, Ishibe-Murakami S, Yamazoe S, Chen JK, et al. ORC1 BAH domain links H4K20 to DNA replication licensing and Meier-Gorlin syndome. Natl Inst Heal. 2012;484(7392):115–9.

133. Botuyan MV, Lee J, Ward IM, Kim JE, Thompson JR, Chen J, et al. Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. Cell. 2006;127:1361–73.

134. Cypress BK. The role of ambulatory medical care in hypertension screening. Am J Public Health. 1979;69(1):19–24.

135. Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet. 2005;37:391–400.

136. Ye J, Ai X, Eugeni EE, Zhang L, Carpenter LR, Mary A, et al. Histone H4 lysine 91 acetylation a core domain modification associated with chromatin assembly. NIH Public Access. 2010;18(1):123–30.

Annex 1 – Modifications of histones

As mentioned in the introduction, there is a wide array of known histone modifications, including acetylation, methylation, phosphorylation and ubiquitylation. These modifications are added to and/or removed from histone proteins by diverse families of proteins including histone acetyltransferases/deacetylases, histone methyltransferases/demethylases, histone kinases/phosphatases, and ubiquitin ligases. Below is described a little about each modification:

- Histone Acetylation

Histone acetylation is a dynamic process. Histone acetyltransferases and histone deacetylases are the enzymes that are responsible for the addition and removal of this modification, and which target particular lysine residues (51).

One of the more heavily studied histone deacetylases is Sir2, a NAD dependant histone deacetylase. It has been shown to be important for the maintenance of silent chromatin and has been implicated as a major regulator of cell aging (51).

- Histone Methylation

One of the most well characterized histone modifications is histone methylation, in addition histones can be methylated on both lysine and arginine residues.

Unlike histone acetylation, histone methylation does not affect the charge of the modified amino acid. In addition, each lysine is able to accept up to three methyl groups, resulting in mono-, di-, and trimethylated lysine forms (52). In many cases, each of these methyl-lysine forms are detectable in the cell at any given time, and have distinct localization patterns, suggesting that each modified form of lysine has a distinct biological role (51–53).

Furthermore, unlike histone acetylation, in which the presence or absence of the modification contributes to an epigenetic state, the location of the methyl mark on the histone dictates its function.

The main family of enzymes that catalyse histone lysine methylation contain the conserved SET domain. Misregulation of these proteins, through overexpression, deletion or chromosomal translocations has been implicated in many types of human cancers.

- Histone Ubiquitination

Ubiquitin is a protein moiety, which is covalently attached to a protein through a series of enzymatic steps (Fig. 31) (54,55).

There is Polyubiquitination, or the addition of multiple ubiquitin moieties, which is typically a sign of protein degradation that is mediated by the proteosome (54,55).

There is also monoubiquitination, or the addition of only a single ubiquitin, which has been observed on histones H2A and H2B (Fig. 32) and is thought to play a role in signalling (56–59). One mechanism by which monoubiquitylation of histones is thought to exert its effects, is by serving as "wedge" which can open up or disrupt chromatin structure (60).

Furthermore, ubiquitylation of histones is reversible, as ubiquitin proteases that can remove this modification have been identified (58). However, the mechanism by which histone ubiquitylation contributes to epigenetic regulation is still unclear.

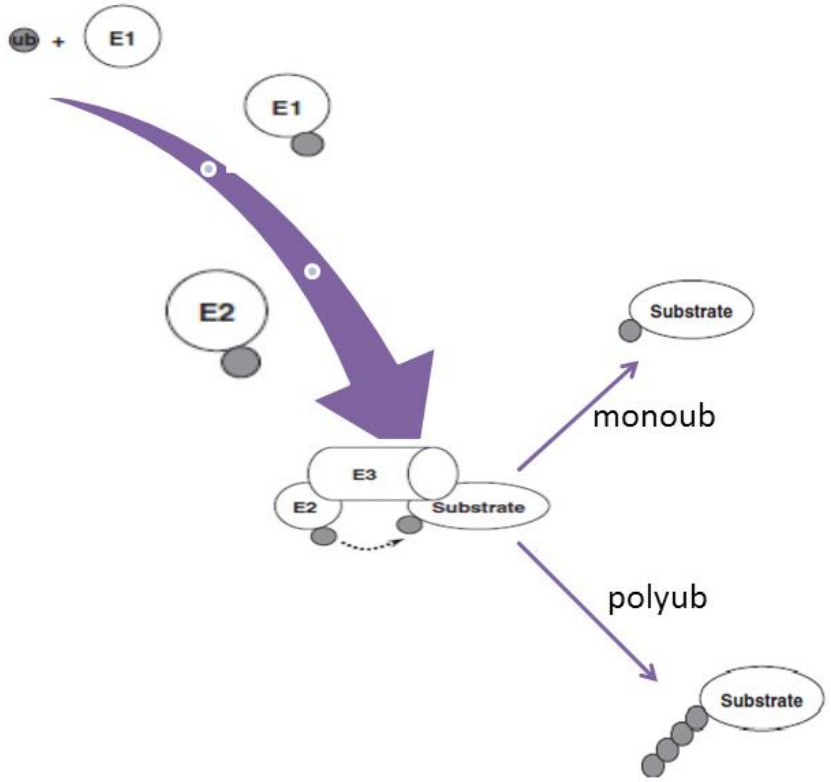


Figure 31: Ubiquitin-conjugating pathway. Ubiquitin is activated by a ubiquitin-activating enzyme (E1) and then transferred to one of the many different ubiquitin-conjugating enzymes (E2).Ubiquitin is conjugated to the correct substrate through the intermediary of a large family of ubiquitin ligases (E3), which bind both the E2-ub complex and substrate. For simplicity, a RING domain E3 is shown. Proteins can be monoubiquitylated, whereby a single ubiquitin is attached, or polyubiquitylated, in which multiple ubiquitin moieties are attached to each other. Adapted image (57)**.**

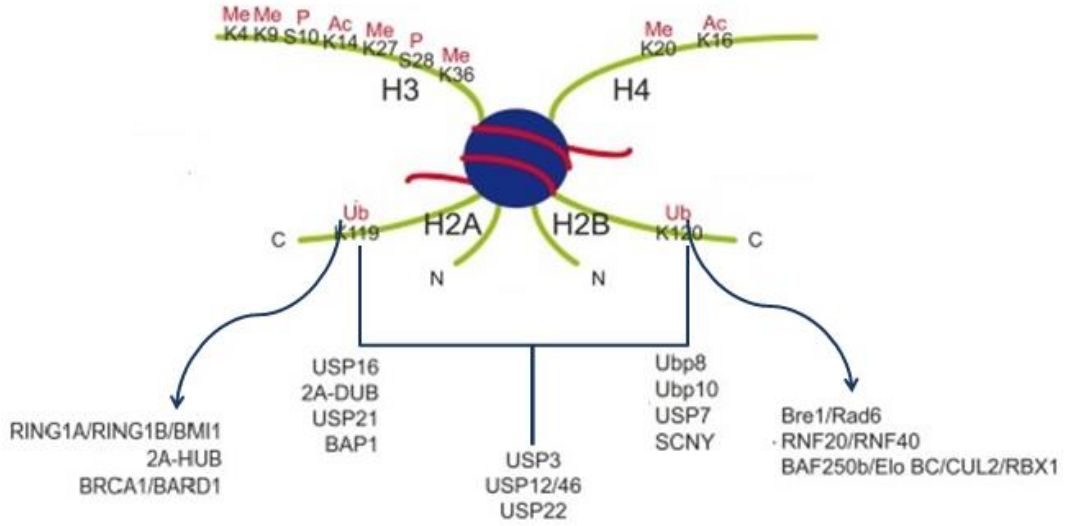


Figure 32: Ubiquitin ligases and deubiquitinating enzymes responsible for monoubiquitination of histones H2A and H2B. Major post-transcriptional modifications on histone H3 and H4 tails are also shown. Ac, acetylation; Me, methylation; P, phosphorylation. Adapted image (61)**.**

- Histone Phosphorylation

Phosphorylation of histone proteins is another post-translational modification that has been studied (17,23). Histone phosphorylation has been linked to a variety of cellular processes, such as chromosome condensation and segregation, activation of transcription, gene silencing, apoptosis and DNA damage repair. Phosphorylation can serve as an activating signal on a target protein, as a transcription factor, to stimulate a response to the external signal resulting in the up- or downregulation of specific genes (62).

We have the example of phosphorylation on histone H3 that has been specifically implicated in cell cycle progression and regulation of gene expression. In addition, phosphorylation of histone proteins may alter chromatin structure by affecting the charge of the histone proteins (63).

A study on phosphorylation of histone H3, observed that, even though phosphorylation of certain histone H3 residues appears to have been conserved, their functional significance might differ among eukaryotes, in particular between metazoan and plants. It was observed that, in mammalian cells, histone H3 is phosphorylated at several sites during mitosis, including serines 10 and 28 (H3S10ph and H3S28ph) and threonines 3 and 11 (H3T3ph and H3T11ph) (20).

Annex 2 – The Three main categories of small ncRNAs

There are three main categories:

a) Short interfering RNAs (siRNA)

siRNAs is the most broadly distributed in both phylogenetic and physiological terms and are characterized by the double-stranded nature of their precursors, such as miRNAs (21,64).

b) MicroRNAs (miRNA)

miRNAs are another class of small RNAs that play a role in gene regulation, which control gene expression acting post-transcriptionally by destabilization or translational repression of the mRNA, inhibiting protein synthesis (22,64). Like as siRNAs, these are also broadly distributed in both phylogenetic and physiological terms and are characterized by the double-stranded nature of their precursors (21,64).

The degree of complementarity between the miRNA and the target mRNA can dictate whether the target mRNA is degraded or translationally blocked. Strong pairing with the mRNA target stimulates cleavage by endonuclease, while weaker pairing can interfere with translation and direct degradation of mRNA. So, varying complementarity means that a particular miRNA has the potential to affect multiple targets (21,22). Presently, it is known that miRNAs can behave as oncogenes and tumour suppressor genes because it has been observed that certain miRNAs are upregulated in certain tumour types, and others play critical roles in cell cycle progression (65).

c) Piwi interacting RNAs (piRNA)

The third class of small RNAs is called piRNAs and unlike the siRNA and miRNA, these are primarily found in animals, exert their functions most clearly in the germline, and derive from precursors which, while being poorly understood, appear to be single stranded (21,22).

Annex 3 – The various tissues present in Consortium

Cell Type	Description
Adrenal – Fetal ▪ Fetal, Adrenal Gland	"The fetal adrenal gland, during most of time the gestation, is almost exclusively dedicated to the production of dehydroepiandrosterone sulfate (DHEA-S)". This specialized ability is unique to primates and occurs because of a specialized fetal zone that composes the bulk of the fetal adrenal gland. The most of the steroid released by the fetal zone is DHEA-S, which is used by the placenta to produce estrogens (66,67).
Brain ▪ Brain Anterior Caudate ▪ Brain Cingulate Gyrus ▪ Brain Hippocampus Middle ▪ Brain Angular Gyrus ▪ Brain Inferior Temporal Lobe ▪ Brain Germinal Matrix ▪ Brain Mid Frontal Lobe ▪ Brain Substantia Nigra ▪ Neurosphere Cultured Cells Cortex Derived ▪ Neurosphere Cultured Cells Ganglionic Eminence Derived ▪ Brain Cerebellum	The brain is composed of many specialized areas that work together: cerebellum is at the base and the back of the brain; basal ganglia are a cluster of structures in the centre of the brain; brain stem is between the spinal cord and the rest of the brain; and cortex is the outermost layer of brain cells (68,69). The brain is also divided into many lobes: frontal lobes; parietal lobes; temporal lobes; occipital lobes (69). The brain is surrounded by a layer of tissue called the meninges (69).
Brain – Fetal ▪ Fetal, Brain ▪ Fetal Spinal Cord	One of the very first systems to develop is fetal nervous system, brain and spinal cord (69).
Breast	The breast is the tissue overlying the chest

▪ Breast Stem Cells ▪ Breast Myoepithelial Cells ▪ Breast Luminal Epithelial Cells ▪ Breast vHMEC	(pectoral) muscles. Women's breasts are constituted of specialized tissue that produces milk (glandular tissue) and still, fatty tissue (69). "The milk-producing part of the breast is organized into 15 to 20 sections, named lobes. Within each lobe are smaller structures, called lobules, where milk is produced. The milk travels through a network of tiny tubes, the ducts, these connect and come together into larger ducts, which eventually exit the skin in the nipple. The dark area of skin surrounding the nipple is called the areola" (69). In addition, the breast is also constituted by connective tissue and ligaments provide support to the breast and give it its shape; nerves provide sensation to the breast; blood vessels, lymph vessels, and lymph nodes (69).
ES Cells ▪ H1 ▪ H9 ▪ HUES1 ▪ HUES3 ▪ HUES6 ▪ HUES8 ▪ HUES9 ▪ HUES13 ▪ HUES28 ▪ HUES44 ▪ HUES45 ▪ HUES48	The Stem Cells are characterized by the capacity for self- renewal and the ability to differentiate into diverse specialized cell types, such as embryonic stem cells, Cancer Stem Cells and Induced Pluripotent Stem cells (IPSCs) (70). Stem cells are the only cells capable of undergoing self-renewal divisions, these divisions are asymmetric in which a stem cell is able to produce an exact copy of itself as well as a daughter cell that undergoes differentiation into the lineages found in differentiated tissues. During

▪ HUES49 ▪ HUES53 ▪ HUES62 ▪ HUES63 ▪ HUES64 ▪ HUES65 ▪ HUES66 ▪ ES-I3 Cell Line ▪ ES-WA7 Cell Line ▪ UCSF-4Star	stem cell expansion and tumorigenesis, stem cells may undergo symmetric divisions in which stem cells produce two different cell populations, one population retains the self-renewing properties of the parental cancer stem cell and the other population is tumor cell with ability to differentiate but without the ability to initiate tumor growth (71).
ES – Derived Cells ▪ H1 Derived Embryoid Body Cultured Cells ▪ H9 Derived Embryoid Body Cultured Cells ▪ HUES1 Derived Embryoid Body Cultured Cells ▪ HUES3 Derived Embryoid Body Cultured Cells ▪ HUES6 Derived Embryoid Body Cultured Cells ▪ HUES45 Derived Embryoid Body Cultured Cells ▪ H1 Derived Neuronal Progenitor Cultured Cells ▪ H9 Derived Neuronal Progenitor Cultured Cells ▪ H9 Derived Neuron Cultured Cells ▪ H1-BMP4 ▪ hESH1 derived mesenchymal ▪ hESH1 derived mesendoderm ▪ hESC Derived CD184+ Endoderm	The Stem Cells derived from other cells, such as embryoid body cultured cells, neuronal progenitor cultured cells, neuron cultured cells, mesenchymal, mesendoderm, CD184+ endoderm cultured cells , CD56+ mesoderm and CD56+ ectoderm cultured cells.

Cultured Cells ▪ hESC Derived CD56+ Mesoderm ▪ hESC Derived CD56+ Ectoderm Cultured Cells	
Exocrine – Endocrine ▪ Adrenal Gland ▪ Pancreas ▪ Spleen ▪ Thymus	The endocrine gland is that secretes its hormones directly into the bloodstream from where it is transported to the target cells, tissues or organs to bring about its effects. However, the exocrine gland is that secretes its hormones into a system of ducts that lead to the external environment (68,69).
Fat – Adult ▪ Mesenchymal Stem Cell Derived Adipocyte Cultured Cells ▪ Adipose Nuclei ▪ Adipose Derived Mesenchymal Stem Cell Cultured Cells ▪ Adipose Tissue	"Mesenchymal stem cells have an intrinsic capacity to differentiate into various cell types in culture or after transplantation". "Adipose cell are connective-tissue cell specialized to synthesize and contain large globules of fat. There are two types of adipose cells: white adipose cells contain large fat droplets, only a small amount of cytoplasm, and flattened, noncentrally located nuclei; and brown adipose cells contain fat droplets of differing size, a large amount of cytoplasm, numerous mitochondria, and round, centrally located nuclei" (68).
GI – Adult ▪ Liver, Adult ▪ Stomach Smooth Muscle ▪ Duodenum Mucosa ▪ Duodenum Smooth Muscle ▪ Pancreatic Islets	The digestive tract is divided into upper gastrointestinal tract (mouth, throat, esophagus and stomach) and lower gastrointestinal tract (small intestine - duodenum, jejunum and ileum, and large intestine - cecum, colon, sigmoid colon,

▪ Colonic Mucosa ▪ Rectal Mucosa ▪ Rectal Smooth Muscle ▪ Colon Smooth Muscle ▪ Esophagus ▪ Gastric ▪ Sigmoid Colon ▪ Small Intestine	rectum and anus) (68,69).
GI – Fetal ▪ Fetal, Intestine Large ▪ Fetal, Intestine Small ▪ Fetal, Stomach	The gastrointestinal tract arises initially during the process of gastrulation from the endoderm of the trilaminar embryo (week 3) and extends from the buccopharyngeal membrane to the cloacal membrane. The tract and associated organs later have contributions from all the germ cell layers (68,69).
GU – Adult ▪ Kidney, Adult ▪ Bladder	Urogenital system or also called genitourinary system, are the organs involved in reproduction and urinary excretion. Although their functions are independent, the structures involved in the excretion and reproduction are morphologically related (68). Thus, the main structures of the urinary system are the kidneys, ureters, bladder and urethra; and the main structures of the reproductive system are the testicles, the vas deferens, urethra and penis in men's case; and are the ovaries, fallopian tubes, uterus and vagina in the female case (68,69).
Heart – Adult ▪ Aorta ▪ Heart	"The heart is a muscular organ, located just behind and slightly left of the breastbone. The heart pumps blood

▪ Left Ventricle ▪ Right Atrium ▪ Right Ventricle	through the network of arteries and veins called the cardiovascular system" (69). "The heart has four chambers: right atrium receives blood from the veins and pumps it to the right ventricle; right ventricle receives blood from the right atrium and pumps it to the lungs, where it is loaded with oxygen; left atrium receives oxygenated blood from the lungs and pumps it to the left ventricle; and left ventricle pumps oxygen-rich blood to the rest of the body. The left ventricle's vigorous contractions create our blood pressure" (68,69). The coronary arteries run along the surface of the heart and provide oxygen-rich blood to the heart muscle. A web of nerve tissue also runs through the heart, conducting the complex signals that govern contraction and relaxation (68). Surrounding the heart is a sac called the pericardium (68).
Heart – Fetal ▪ Fetal, Heart	The cardiovascular system is the first to operate in the embryo, mainly due to the need for an efficient method of oxygen and nutrient uptake. The primitive heart is composed of venous sinus, sinoatrial valve, primitive atrium, primitive ventricle, bulbus and truncus arteriosus, and between the middle of the 4th week and the end of the 5th week of pregnancy occurs septation of the primitive heart (68,69). That is, the septation of the atrioventricular canal.

Hematopoietic Stem ▪ CD34, Primary Cells ▪ CD34, Mobilized Primary Cells ▪ CD34, Cultured Cells	"The stem cells that form blood and immune cells are known as hematopoietic stem cells. These are ultimately responsible for the constant renewal of blood—the production of billions of new blood cells each day". So, hematopoietic stem cell is a cell isolated from the blood or bone marrow that can renew itself, can differentiate to a variety of specialized cells, can mobilize out of the bone marrow into circulating blood, and can undergo programmed cell death, called apoptosis—a process by which cells that are detrimental or unneeded self-destruct (72).
IPs Cells ▪ iPS 4.7 ▪ iPS 6.9 ▪ iPS-11a ▪ iPS-11b ▪ iPS-11c ▪ iPS-15b ▪ iPS-17a ▪ iPS-17b ▪ iPS-18a ▪ iPS-18b ▪ iPS-18c ▪ iPS-19.7 ▪ iPS-19.11 ▪ iPS-20b ▪ iPS-27b ▪ iPS-27e	" IPSCs are adult cells that have been genetically reprogrammed to an embryonic stem cell–like state by being forced to express genes and factors important for maintaining the defining properties of embryonic stem cells" (73,74). IPSCs demonstrate important characteristics of pluripotent stem cells, including expressing stem cell markers, forming tumours containing cells from all three germ layers, and being able to contribute to many different tissues when injected into mouse embryos at a very early stage in development, and still express stem cell markers and are capable of generating cells characteristic of all three germ layers (73,74). These cells are already useful tools for drug development and modelling of

	diseases, and scientists hope to use them in transplantation medicine (73,74).
Kidney – Fetal ▪ Fetal, Kidney ▪ Fetal, Kidney Left ▪ Fetal, Kidney Right ▪ Fetal, Renal Cortex ▪ Fetal, Renal Cortex Left ▪ Fetal, Renal Cortex Right ▪ Fetal, Renal Pelvis ▪ Fetal, Renal Pelvis Left ▪ Fetal, Renal Pelvis Right	Kidney is the main organ of the urinary tract. It is a kind of factory to eliminate blood unnecessary or harmful substances, while losing the substances that are important for the functioning of the body, where the glomeruli (68,69). These glomeruli filter water and various substances.
Lung – Adult ▪ Lung	The lung is part of the respiratory system, which is an expert system for the exchange of gases. Its main function is to oxygenate the blood and remove carbon dioxide or carbon dioxide from the body. The human being has two lungs located in the chest cavity, covered by a protective membrane known as the pleura. Each lung is divided into lobes, the right lung is larger and has three lobes; the left has only two (69,75).
Lung – Fetal ▪ Fetal, Lung ▪ Fetal, Lung Left ▪ Fetal, Lung Right ▪ Fetal, Lung Fibroblast (IMR90)	The development of the fetal lung is done in four stages, ranging from five weeks of pregnancy to early childhood, school age. The first step corresponds to the processes pseudoglandular period and from 5 to 16 weeks of gestation. After this period, only the structures involved in gas exchange will not be formed. Fetal breathing is not yet possible. The second stage corresponds to the canalicular period, and occurs between 16 and 24 weeks of gestation. At

	the end of this step the structures responsible for gas exchange, although immature, are already formed. Breathing is already possible, although limited. The third stage corresponds to the vesicular period and takes place from 24 weeks gestation until birth. The fourth stage corresponds to the alveolar period, and occurs from birth to school age (69,75).
Muscle - Adult ▪ Skeletal Muscle ▪ Muscle Satellite Cultured Cells ▪ Psoas Muscle	Our walking ability depends on the combined action of bones, joints and muscle, under the regulation of the nervous system. The muscular system is formed by the body of the set of muscles, and we have big muscles such as the thigh, and small muscles, such as certain muscles of the face. They may be rounded (the orbicular eye, for example); plans (the skull, etc.); or spindle (like arms). In general, there are three types of muscles: non-striated muscle (smooth muscle); Skeletal muscle; Cardiac striated muscle (68,69).
Muscle - Fetal ▪ Fetal, Muscle Arm ▪ Fetal, Muscle Back ▪ Fetal, Muscle Leg ▪ Fetal, Muscle Lower Limb ▪ Fetal, Muscle Upper Limb ▪ Fetal, Muscle Trunk ▪ Fetal, Muscle Upper Trunk	The Muscular System develops from mesoderm except the iris muscles, which originates from neuroectoderma. Since the myoblasts, embryonic muscle cells, mesenchymal stem (embryonic connective tissue) (68,69).
Placenta – Fetal ▪ Fetal, Placenta	The placenta is an organ that exists only during pregnancy and has several

▪ Placenta Amnion ▪ Placenta Basal Plate ▪ Placenta Chorion Smooth ▪ Placenta Trophoblast Primary Cells ▪ Placenta Villi	functions such as: Provide nutrients and oxygen to the baby; Production of hormones; Baby immune protection; Baby impact protection in the womb; It also eliminates the waste that the baby produces, such as urine (68,69).
Reproductive – Adult ▪ Ovary ▪ Testis Spermatozoa Primary Cells	There is the female reproductive system constituted by the ovaries, fallopian tubes, uterus and vagina; and the male reproductive system constituted by the scrotum, testicles, epididymis, vas deferens, urethra, seminal vesicles, prostate and penis (68,69).
Reproductive – Fetal ▪ Fetal Ovary	The ovary makes part of the female reproductive system (68).
Skin – Fetal ▪ Fibroblasts Fetal Skin Abdomen ▪ Fibroblasts Fetal Skin Biceps Left ▪ Fibroblasts Fetal Skin Biceps Right ▪ Fibroblasts Fetal Skin Quadriceps Left ▪ Fibroblasts Fetal Skin Quadriceps Right ▪ Fibroblasts Fetal Skin Upper Back ▪ Fibroblasts Fetal Skin Upper Back	The skin is the body that involves the body determining their limit with the external environment and performs various functions, such as thermal regulation, organic defense, blood flow control, protection against various agents of the environment and sensory functions (heat, cold, pressure, pain and touch). It is composed of three layers: epidermis, dermis and hypodermis, the outermost to the deepest, respectively (76).
Spleen – Fetal ▪ Fetal, Spleen ▪ Fibroblasts Fetal Skin Back	"In the fetus, the spleen acts as a hemopoietic center until late in gestation, and produces lymphocytes and monocytes

	throughout life. The spleen arises as an aggregation of reticular mesenchymal cells in the dorsal mesentery of the stomach during the sixth to seventh weeks, menstrual age. It acquires it is characteristic crescent shape early in the fetal period" (69).
Stromal – Connective ▪ Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells ▪ Chondrocytes from Bone Marrow Derived Mesenchymal Stem Cell Cultured Cells ▪ Primary Fibroblast ▪ Fetal Skin ▪ Foreskin Fibroblast Primary Cells ▪ Foreskin Keratinocyte Primary Cells ▪ Foreskin Melanocyte Primary Cells ▪ Breast Fibroblast Primary Cells	Connective Tissue Proper is a group of connective tissues designed generally to bind parenchymal tissues, blood vessels and nerves together to form organs and to passively transfer mechanical tension from point to point within the organ or within the animal body. The stromal connective tissues belong to this group, that are the common connective tissues found in the stroma, namely loose connective tissue and dense irregular connective tissue (76).
Thymus – Fetal ▪ Fetal, Thymus	The thymus gland is an organ responsible for immune functions lymphoepithelial both the uterus and out of the uterus, which originates embryologically from the third branchial pouch in the fetus, which appears fully developed in the 3rd. month of pregnancy and is composed of two lobes joined by a connective tissue (77). The thymus grows regularly during the period of childhood, reaching its peak in

	the early stages of puberty and begins to atrophy in the early juvenile period being replaced by adipose tissue (69,77).
White Blood ▪ CD3+ Total Unmobilized ▪ CD3+ Total Mobilized ▪ CD8, Primary Cells ▪ CD8, Naive Primary Cells ▪ Mobilized CD8 Primary Cells ▪ CD4, Primary Cells ▪ CD4, Mobilized Primary Cells ▪ CD4, Memory Primary Cells ▪ CD4+ CD25- CD45RA+ Naive Primary Cells ▪ CD4, Naive Primary Cells ▪ CD4+ CD25+ CD127- Treg Primary Cells ▪ Treg Primary Cells ▪ Th17 Primary Cells ▪ CD4+ CD25- IL17+ PMA-Ionomcyin stimulated Th17 Primary ▪ CD4+ CD25- CD45RO+ Memory Primary Cells ▪ CD4+ CD25- IL17- PMA-Ionomycin stimulated MACS purified Th Primary Cells ▪ CD4+ CD25- Th Primary Cells ▪ CD4+ CD25int CD127+ Tmem Primary Cells	"White blood cells are also called leukocytes, these are an important part of the immune system and originate in the bone marrow, but circulate throughout the bloodstream. Its function is help fight infections by attacking bacteria, viruses, and germs that invade the body. There are five major types of white blood cells: neutrophils; lymphocytes; eosinophils; monocytes; basophils" (72).

▪ CD14, Primary Cells ▪ CD56, Primary Cells ▪ CD8, Mobilized Cells ▪ CD15, Primary Cells ▪ CD19, Primary Cells ▪ CD20, Primary Cells ▪ CD56, Mobilized Cells ▪ Peripheral Blood Mononuclear Primary Cells ▪ CD4+ CD25- CD45RA+ Naive Primary Cells ▪ CD4+ CD25- CD45RO+ Memory Primary Cells	

Annex 4 – Characteristics of each histone modification

Histone	Description
H2A.Z	"Between core histones, the H2A family exhibits highest sequence divergence, resulting in the largest number of variants known". Being that, H2A variants differ mostly in their C-terminus. Consequently, H2A variant incorporation has the potential to strongly regulate DNA organization on several levels resulting in meaningful biological output" (78). "The H2A family, on the other hand, contains a plethora of variants with some 'universal variants' found in almost all organisms, namely H2A.Z and H2A.X" (78). Histone H2A.X was, together with H2A.Z, first described in 1980 (78). "After DNA damage, this serine becomes phosphorylated and renders H2A.X an important player in preserving genome integrity. Thus, H2A.X has been shown to be involved in the DNA damage response." "Histone H2A.Z is an almost universal variant, which evolved early and suggest that H2A.Z fulfils specific and unique functions that cannot be carried out by other H2A variants. Like other histones, H2A.Z can be post-translationally modified by: H2A.Z sumoylation has been implicated in DNA repair in S. cerevisiae (79); ubiquitination correlates with localization to the inactive X chromosome in mammals (80), whereas N-terminal acetylation leads to nucleosome destabilization (81); H2A.Z acetylation works as a switch-like mechanism to modulate H2A.Z nucleosome stability, ascribing repressive functions to the unmodified and activating functions to the acetylated form. Furthermore, acetylated H2A.Z was found associated with transcription, revealing that this variant of H2 is enriched at gene promoters. It has been found that H2A.Z can have both activating and repressive influences on transcription" (78). "The biological function of H2A.Z has been extensively studied revealing roles in transcription regulation, DNA repair, heterochromatin formation, chromosome segregation and mitosis. Because of space

	constraints, we cannot discuss all aspects of H2A.Z biology" (78,82).
H2AK5ac	Histone H2A Lys5 is preferentially acetylated by Tip60 and deacetylated by HDAC3, that causes its release from damaged chromatin and activates DNA damage response (83).
H2AK9ac	"All acetylations are correlated with gene expression, consistent with their involvement in transcriptional activation". However, different acetylations may target different regions of genes. For example, H2AK9ac, H2BK5ac, H3K9ac, H3K18ac, H3K27ac, H3K36ac and H4K91ac are mainly located in the region surrounding the transcription start sites, whereas H2BK12ac, H2BK20ac, H2BK120ac, H3K4ac, H4K5ac, H4K8ac, H4K12ac and H4K16ac are elevated in the promoter and transcribed regions of active genes (84).
H2BK5ac	H2BK5 acetylation shows also a robust peak at the transcription start site of active (84).
H2BK12ac	"H2BK12 is shown to be heavily deacetylated on exposure to Nickel(II) which is supposed to be carcinogenic. This also show to be enriched at the transcription start site" (84). H2BK12 acetylation is found at differentially methylated regions of imprinted genes (85).
H2BK15ac	Acetylated H2BK15 levels are reported to rise after RGC-32 knockdown in colon cancer cell lines (86).
H2BK20ac	"H2BK20 is shown to be heavily deacetylated on exposure to Nickel(II) which is supposed to be carcinogenic" (87). This also show to be enriched at the transcription start site (84).
H2BK120ac	This another histone that show to be enriched at the transcription start site (84).
H3K4ac	This histone show also to be enriched at the transcription start site (84).
H3K4me1	H3K4me1 and H3K27ac are histones that flank active enhancers. The enhancers that are not yet active but that are primed for activation either at a later developmental time point or in response to external stimuli that can be pre-marked by H3K4me1. "Lastly, latent enhancers are located in closed chromatin and are not pre-marked by known histone modifications, but in the presence of external stimuli the DNA becomes

	accessible, and flanking nucleosomes acquire H3K4me1 and H3K27ac marks" (88).
H3K4me2	Dimethylation of histone H3K4 is linked to active transcription (89). It is associated with lung cancer (90); prostate cancer (91); Kidney cancer (91); neurological disorders: Fragile X syndrome (92); colon cancer (93).
H3K4me3	H3K4 is associated with activation. H3K4me3 recruits the chromatin remodeling factors CHD1 (94) and BPTF (95) which open chromatin, while preventing the binding of the repressive NuRD (96) and INHAT complexes (97). H3K4 methylation enzymes were initially identified as regulators of Hox genes. So, a relatively recent paper has further elucidated the mechanism by which H3K4me3 promotes rapid gene activation (98).
H3K9ac	H3K9 can turn genes on by getting acetylated, so how can silence them just as easily when methylated. H3K9ac is a particularly important acetylation because it is highly correlated with active promoters. H3K9ac also has a high co-occurrence with H3K14ac and H3K4me3 which together are these three marks are the hallmark of active gene promoters (99).
H3K9me1	There are, mono, di, and tri H3K9 methylations, being that H3K9me1 is enriched at the transcriptional start site of active genes (95), and H3K9me2/3 were both found more often at silenced genes. H3K9 methylation is linked to transcriptional repression. As is known LSD1 dependent demethylation of H3K9 leads to de-repression of androgen receptor target genes (100). The monomethylated H4K20 and H3K9 act cooperatively to mark distinct regions of silent chromatin within the mammalian epigenome (101).
H3K9me3	"H3K9 methylation is the mark of heterochromatin, being that, heterochromatin is the condensed, transcriptionally inactive state of chromatin. It can be facultative or constitutive. H3K9me3 binds heterochromatin protein 1 (HP1) to constitutive heterochromatin (96). Knowing that HP1 is responsible for transcriptional repression and the

	actual formation and maintenance of heterochromatin. HP1 also recruits DNA methyltransferase 3b, providing one of the best examples of the interplay between histone methylation and DNA methylation". H3K9me3 as H3K27me3 are known for their roles associated to repression in the genic and nongenic regions of metazoan genomes. Many complexes are known to be responsible for generating these marks, such as polycomb group complexes and H3K9 methylases (102).
H3K14ac	"Histone H3K14 acetylation is critical for the recruitment of TFIID at the IFN-gamma locus, so it is important for eliciting proper immune response" (103).
H3K18ac	Studies show a significant correlation between low expression of H3K18ac and better prognosis of patients with esophageal squamous cell carcinoma (104). H3K18 acetylation shows a robust peak at the transcription start site of active and poised genes (84). It is associated with prostate cancer (105); lung cancer (91); kidney cancer (91); brain tumor (glioblastoma) (105).
H3K23ac	H3K23 acetylation and H3K4 methylation are part of a non-canonical histone signature that is recognised by chromatin regulator tripartite motif-containing 24 (TRIM24) which binds chromatin and oestrogen receptor to activate the oestrogen-dependent genes associated with cellular proliferation and tumour development. The aberrant expression of TRIM24 negatively correlates with survival of breast cancer patients (106).
H3K23me2	H3K23 is extensively modified by methylation and that H3K9me3 is exclusively detected on histone tails with H3K23me2. The histone H3K23me2 is enriched in heterochromatic regions. The chromatin immunoprecipitation approaches revealed a positive correlation between H3K23me2 and repressive marks, and by immunofluorescence analyses, H3K23me2 appears differentially regulated in germ and somatic cells, in part by the action of the histone demethylase JMJD-1.2. Thus, This defines repressive domains and contributes to organizing the genome in distinct heterochromatic regions

	during embryogenesis (107).
H3K27ac	Since a lysine residue cannot be both methylated and acetylated you would expect H3K27ac to be antagonistic to the repression of gene expression by H3K27me2/3. Indeed, data are showing that H3K27ac is associated with active transcription and antagonism of H3K27me3 regulated genes (108).
H3K27me3	H3K27 is known for shutting down transcription. H3K27 is associated with inactive gene promoters when it is trimethylated. This histone acts in opposition to H3K4me3. As is already known, the most histone methylations are catalysed by many enzymes but H3K27me3 has only one known methyltransferase: EZH2 (109). "EZH2 is part of the PRC2 complex which is responsible for the repression many genes involved in development and cell differentiation" (110,111). So, it is believed that H3K27me3 is critical for the repression of developmental genes and she is also an important mark of the inactive X chromosome (112) .
H3K36me3	H3K36 plays a role in transcriptional activation (113). "The histone may have a change mono, di, and trimethylation, and this states differ from each other in their distributions and functional roles. About the role of H3K36me1 remains unclear; however, H3K36me2 is relatively well characterized; for example, H3K36me2 has a role in double-strand break repair. H3K36me2 is deposited near the double-strand breaks early, and then serves to recruit early repair factors such as NBS1 and Ku70" (114). H3K36me3 is involved in defining exons. Exons are enriched in nucleosomes in general, but these nucleosomes are also enriched in certain histone modifications including H3K79, H4K20, and especially H3K36me3. It is believed that this pattern influences alternative splicing in some way, perhaps by signalling effector proteins to mark particular exons for inclusion in the final transcript (115).
H3K56ac	Histone H3 acetylated at Lysine 56 is assembled into chromatin in human cells, forming foci that colocalise with sites of DNA repair. Acetylation this histone is increased in multiple types of cancer,

	correlating with increased levels of the histone chaperone ASF1A, in these tumours. “ASF1A is required for the acetylation and CAF-1 is required for the incorporation of acetylated H3 into the nucleosome” (116). “H3K56ac levels in human cells are differentially regulated at telomeres and globally in response to cell cycle arrest” (117). During DNA damage, H3K56 acetylation levels increased, as this acetylated H3K56 is also localised at the DNA repair sites. It also colocalised with other proteins involved in DNA damage signalling pathways including phospho-ATM, Chk2, and p53 argue its involvement in DNA damage repair (118).
H3K79me1	“The methylation of histone H3 K 79 was the first modification identified in the globular core domain of the histones”. Studies suggest that specific H3K79 methylation states play distinct roles in the response to UV-induced DNA damage, so H3K79me is associated with DNA double-strand break (DSB) responses (119).
H3K79me2	“H3K79 dimethylation is cell cycle dependent”. This histone level decreases during S phase, reaches its lowest level in G2, increases during M phase, and keeps at a high level during G1 phase (120,121).
H3T11ph	H3T11 phosphorylation is a mitosis specific modification (122). It is phosphorylated at androgen receptor gene targets by PRN1 which acts as a co-activator of androgen receptor dependent transcription. Besides that promotes demethylation of histone H3 'Lys-9' (H3K9me) by JMJD2C. The levels of PRK1 and phosphorylated H3T11 correlate with Gleason scores of prostate carcinomas (123).
H4K5ac	“H4K5 is the closest lysine residue to the N-terminal tail of histone H4 and the histone H4 forms a strong tetramer with histone H3. As histone H3, H4 has a long N-terminal tail that is subject to various acetylations and methylations that are associated with many cellular processes” (124). “H4K5 is acetylation, and it is implicated in epigenetic bookmarking, in other words, Epigenetic bookmarking the name given to a proposed process that allows gene expression patterns to be faithfully passed to

	daughter cells through mitosis. Important cell-type specific genes are marked in some way that prevents them from being compacted during mitosis and ensures their rapid transcription. H4K5ac has been implicated as one such mark" (125). "Zhao et al. (2011) found that transcriptional activity during interphase causes acetylation of H4K5 and H4K12 which are passed through mitosis. These marks then recruit BRD4 which de-compacts the local chromatin environment and permits transcriptional activation. Similarly, other papers have found that H4K5ac can serve as a primer for rapid transcription in other contexts, such as, H4K5ac appears to prime activity-dependent genes expressed during learning (126). The mark may permit the rapid expression of certain genes required during the learning process".
H4K8ac	H4K8 is another lysine on that tail of histone H4 and it is only known to be acetylated. This group of lysines are known to act as transcriptional activators (84). This proposes that H4K8ac serves to facilitate transcriptional elongation (127). Wang et al. (2008), found that H4K8ac is part of 17 modifications that occupy most active promoters. This study also found that H4K8ac was found more often in active promoters and transcribed regions than the others groups which were found more at transcriptional start sites. H4K8ac is catalyzed and read by a slightly different group of enzymes than other H4 lysines, mMany acetyltransferases catalyze H4K8ac (128).
H4K12ac	Like H4K8ac, H4K12ac is part of a "backbone" of histone modifications that are associated with active promoters (84). H4K12ac is localized to the promoter, like other H4 acetylations (84). H4K12ac appears to be vital for learning and memory. One study showed its importance in controlling gene expression in the hippocampus associated with memory consolidation, and showed also that normal age-related decline in memory could be ameliorated by restoration of H4K12ac (127). "Recently it has been shown that H4K12ac is important in paternal

	influence on early gene expression in the zygote. Sperm chromatin is packaged with protamines instead of histones for the majority (85%) of the genome, serving to inactivate transcription. The rest of the genome is occupied by the traditional histone octamers. Data have shown some of this remaining portion of the chromatin bear H4 lysine acetylations that are passed into the zygote and can be involved in early embryo development" (129). Other studies found that H4K12ac was enriched surrounding transcriptional start sites, and that the genes it occupied had a significant bias toward developmental functions (130).
H4K20me1	H4K20 is methylated but not acetylated. Like all lysine residues, H4K20 can be mono, di, or tri methylated, and these methylation states have different spatial disruptions and functions. H4K20me1 is associated with transcriptional activation, being that the most highly transcribed group of genes tend to have H4K20me1 present in addition to the core group of modifications at active promoters (79). "H4K20me1 is also important for cell cycle regulation, because PR-Set7 (enzyme that catalyses this modification) levels oscillate during the cell cycle, and regulate chromatin condensation and mitotic progression" (84,131). H4K20me2 is important for cell cycle control, particularly for marking points of origin for DNA replication, and it is also important in the DNA damage response (132,133). "H4K20me3 have a much different function that the other two methylation states. Usually, H4K20me3 is associated with repression of transcription when present at promoters and it is also important for the silencing of repetitive DNA and transposons (134). These results can explain why the loss of H4K20me3 has be identified as a hallmark of cancer. Along with reduction of H4K16ac, loss of H4K20me3 is a near universal characteristic of human cancers" (135).
H4K91ac	"Histone H4 showed the presence of modifications in the globular core domain. One site of acetylation, lysine 91, localizes to a region of H4 that is important for the interaction of the H3/H4 tetramer with

	H2A/H2B dimers" (136). So, the molecular genetics and biochemical evidence indicate that the acetylation of lysine 91 influences the process of chromatin assembly and may function to modulate the formation of histone octamers. Having thus, increased sensitivity to DNA damaging agents (136).

Annex 5 – R scripts

R scripts used in program R throughout this work:

For deliver these outcomes, it was necessary to read a frag-stats file and complete it with oligonucleotide frequency counts of a context around each variation. I am grateful to teacher João Rodrigues of DETI, for providing this function, although we have adapted to our work.

R scripts used:

```
source("read-frag-stats-pos2.R")
```

As our files taken from the database, had multiple columns with different information, we had to create a function to just stay with the columns that interest to get the results, meaning, take the last two columns (metadata), being one optional variable.

R scripts used:

```
colClasses=c(NA, NA, NA,"NULL","NULL")
```

To reading of the bed file may be compressed or not. Where can we read only the first 100 line, for exemple ("Nrow = 100"). And we also have to give an indication to read only the columns that interests us ("ColClasses = colClasses").

Our file changed its name to "t".

R scripts used:

```
t = read.table("File.bed.gz", colClasses=colClasses, nrow=100)
```

The next function is to know how many chromosomes have our file "t".

R scripts used:

```
unique(t)
```

In order to see the top of the table that will work just this function.

R scripts used:

```
head(t)
```

It will be not our case, but if we want to filter out only a chromosome analysis is also possible, for example, chromosome 1.

R scripts used:

```
t=t[t$CHROM=="chr1",]
```

To obtaining the size of the object.

R scripts used:

```
dim(t)
```

The command used for the frequencies of oligonucleotides, has a set of functions. But before applying this command, we had to create a function to exclude the mitochondria because they do not want to include in our analysis.

R scripts used:

```
# Exclude the mitochondria
level=levels(t$CHROM)
level=level[level!="chrM"]
# For nucleotides, for example
k=1
Tc = c(1:4^k)
for (chr in level){
t1=t[t$CHROM==chr,]
tt = oligonucleotideFrequency.frag.neighborhood.pos2(t1, k)
Tc=cbind(Tc,colSums(tt))
}

colnames(Tc) <- c("Tc", level)
```

To get these results in excel, we have to install a package ("Write.xlsx") and create a role, giving the file a name.

R scripts used:

```
Dat ← data.frame(what we save)
write.table(dat, "name.txt")
```

This program can also read files from Excel.

R scripts used:

```
read.xlsx("name.xlsx", 1)
```

After having the first part of the intended results, we want to do statistical analysis. To do this, we use the previous command for read files from Excel, and we name the columns. Thus started doing the statistical analysis:

R scripts used:

```
#Read files from Excel
library("xlsx")
a = read.xlsx("name.xlsx", 1)
# Identify the columns in the program for nucleotides, for example
rownames(a) <-  c("A", "C", "G", "T")
# Or identify the columns in the program for dinucleotide
rownames(a) <-  c("AA", "AC", "AG", "AT", "CA", "CC", "CG", "CT", "GA", "GC", "GG", "GT", "TA", "TC", "TG", "TT")
# Calculate the chi-square test
x = chisq.test(a)
# Calculate the analysis of standardized residuals
p= x$stdres
# Get the two-dimensional representation of data in which values are represented by colors
library(gplots)
my_palette=colorRampPalette(c("red", "orange", "black", "yellow", "green"))(n = 999)
colors=c(seq(-1000,-101,length=200),seq(-100,-11,length=200),seq(-10,10,length=200),seq(11,100,length=200), seq(101,1000,length=200))
heatmap.2(p, col=my_palette, breaks=colors)
# Calculate the measure of association Cramer's V.
library(questionr)
v=cramer.v(a)
```

For the completion of the control we had to complementary. For this we use the package GenomicRanges.

R scripts used:

```
source("http://bioconductor.org/biocLite.R")
biocLite("Biostrings")
library(IRanges)
library(GenomicRanges)

colClasses=c(NA, NA, NA,"NULL","NULL")
t = read.table("union.bed.gz", colClasses=colClasses)
names(t)=c("CHROM","POS1","POS2")              → Put names in columns
B=IRanges(start=t[[2]], end=t[[3]])
grB <- GRanges(t[[1]], B, strand="+")
grB

require(xlsx)
C0=read.xlsx("control0.xlsx", 1)
A=IRanges(start=C0[[2]], end=C0[[3]])
grA <- GRanges(C0[[1]], A, strand="+")
D=setdiff(grB, grA)
grB=setdiff(grB,D)

D=setdiff(grA, grB)
aux=seqnames(D)
aux<- data.frame(aux)
Pos=ranges(D)
Pos<- data.frame(Pos)
control=cbind(aux,Pos)
names(control)=c("CHROM","POS1","POS2")

# Exclude the mitochondria
level=levels(control$CHROM)
level=level[level!="chrM"]


# This function is necessary to read a snp-stats file and complete it with oligonucleotide frequency counts of a context around each variation

source("read-snp-stats-pos2.R")

# The command used for the frequencies of oligonucleotides
k=1 (nucleotide, for exemple)
Tc = c(1:4^k)
for (chr in level){
t1=control[control$CHROM==chr,]
tt = oligonucleotideFrequency.frag.neighborhood.pos2(t1, k)
Tc=cbind(Tc,colSums(tt))
}
colnames(Tc) <- c("Tc", level)
```

Annex 6 – Histone union and histone intersection

		Union				Intersection			
		X^2	df	p-value	Cramer's V	X^2	df	p-value	Cramer's V
H2AZ	Nucleotide	10063000	69	<0.001	0.0243386	79652	69	<0.001	0.01740994
	Dinucleotide	22290000	345	<0.001	0.0161998	214010	345	<0.001	0.01283165
	Trinucleotide	34156000	1449	<0.001	0.01619488	376980	1449	<0.001	0.01382865
	Tetranucleotide	46218000	5865	<0.001	0.01883907	678130	5865	<0.001	0.01864872
H2AK5ac	Nucleotide	8398000	69	<0.001	0.02319443	9629.3	69	<0.001	0.0128653
	Dinucleotide	18742000	345	<0.001	0.01550353	28476	345	<0.001	0.01018692
	Trinucleotide	28852000	1449	<0.001	0.01554196	51427	1449	<0.001	0.01140307
	Tetranucleotide	39203000	5865	<0.001	0.01812571	98694	5865	<0.001	0.01629836
H2AK9ac	Nucleotide	1182900	69	<0.001	0.01774537	173590	69	<0.001	0.01755841
	Dinucleotide	2773900	345	<0.001	0.01227663	422120	345	<0.001	0.01249854
	Trinucleotide	4317600	1449	<0.001	0.01249791	673980	1449	<0.001	0.01302934
	Tetranucleotide	5965200	5865	<0.001	0.01484653	1021600	5865	<0.001	0.01639573
H2BK5ac	Nucleotide	8423500	69	<0.001	0.02330746	3255.7	69	<0.001	0.02197853
	Dinucleotide	18777000	345	<0.001	0.01557157	11069	345	<0.001	0.01870049
	Trinucleotide	28863000	1449	<0.001	0.01560005	23946	1449	<0.001	0.02296858
	Tetranucleotide	39156000	5865	<0.001	0.01818094	59139	5865	<0.001	0.03732939
H2BK12ac	Nucleotide	7360200	69	<0.001	0.02241408	8573.9	69	<0.001	0.01376977
	Dinucleotide	16525000	345	<0.001	0.01502819	27334	345	<0.001	0.01128178
	Trinucleotide	25521000	1449	<0.001	0.01509051	53095	1449	<0.001	0.01304699
	Tetranucleotide	34780000	5865	<0.001	0.01762608	112950	5865	<0.001	0.0195584
H2BK15ac	Nucleotide	8627700	69	<0.001	0.02335652	29197	69	<0.001	0.01503076
	Dinucleotide	19239000	345	<0.001	0.01560198	83473	345	<0.001	0.01146271
	Trinucleotide	29612000	1449	<0.001	0.01563603	152830	1449	<0.001	0.0126342
	Tetranucleotide	40231000	5865	<0.001	0.01823015	290440	5865	<0.001	0.01756811
H2BK20ac	Nucleotide	7304500	69	<0.001	0.02279845	220730	69	<0.001	0.01713053
	Dinucleotide	16375000	345	<0.001	0.01527556	560850	345	<0.001	0.01227545
	Trinucleotide	25219000	1449	<0.001	0.01531888	965860	1449	<0.001	0.01307765
	Tetranucleotide	34255000	5865	<0.001	0.01786481	1607900	5865	<0.001	0.01696204
H2BK120ac	Nucleotide	7805600	69	<0.001	0.02273214	8541.2	69	<0.001	0.02069741
	Dinucleotide	17470000	345	<0.001	0.01521481	36763	345	<0.001	0.01937769
	Trinucleotide	26933000	1449	<0.001	0.01526231	103870	1449	<0.001	0.02654775
	Tetranucleotide	36649000	5865	<0.001	0.01781086	263890	5865	<0.001	0.0427055

H3K4ac	Nucleotide	7785600	69	<0.001	0.02261121	7265.7	69	<0.001	0.01574844
	Dinucleotide	17437000	345	<0.001	0.01513886	22648	345	<0.001	0.01277403
	Trinucleotide	26915000	1449	<0.001	0.01519519	43784	1449	<0.001	0.01475758
	Tetranucleotide	36661000	5865	<0.001	0.01774081	98015	5865	<0.001	0.02272458
H3K4me1	Nucleotide	10061000	69	<0.001	0.02438253	358.9	63	<0.001	0.07653186
	Dinucleotide	22283000	345	<0.001	0.01622801	1339	315	<0.001	0.06700341
	Trinucleotide	34137000	1449	<0.001	0.01622122	3669.1	1323	<0.001	0.0950398
	Tetranucleotide	46181000	5865	<0.001	0.01886729	10443	5355	<0.001	0.1625373
H3K4me2	Nucleotide	9039200	69	<0.001	0.02374319	15963	69	<0.001	0.01212389
	Dinucleotide	20109000	345	<0.001	0.01584191	49321	345	<0.001	0.009621216
	Trinucleotide	30882000	1449	<0.001	0.01585863	83609	1449	<0.001	0.01021452
	Tetranucleotide	41854000	5865	<0.001	0.0184672	139300	5865	<0.001	0.01331299
H3K4me3	Nucleotide	10053000	69	<0.001	0.02437461	606.25	69	<0.001	0.007432931
	Dinucleotide	22266000	345	<0.001	0.01622313	2717.3	345	<0.001	0.007072333
	Trinucleotide	34112000	1449	<0.001	0.0162168	6783.5	1449	<0.001	0.009069115
	Tetranucleotide	46149000	5865	<0.001	0.01886267	18124	5865	<0.001	0.01489811
H3K9ac	Nucleotide	10051000	69	<0.001	0.02443599	3289.1	69	<0.001	0.0188334
	Dinucleotide	22258000	345	<0.001	0.01626303	10720	345	<0.001	0.01550041
	Trinucleotide	34087000	1449	<0.001	0.01625357	22552	1449	<0.001	0.01852131
	Tetranucleotide	46097000	5865	<0.001	0.01890194	54964	5865	<0.001	0.02949854
H3K9me1	Nucleotide	3881400	69	<0.001	0.02149542	860540	69	<0.001	0.01698349
	Dinucleotide	8815500	345	<0.001	0.01450667	2044500	345	<0.001	0.01173788
	Trinucleotide	13628000	1449	<0.001	0.0145853	3225500	1449	<0.001	0.01193756
	Tetranucleotide	18645000	5865	<0.001	0.01708301	4520100	5865	<0.001	0.01416892
H3K9me3	Nucleotide	10042000	69	<0.001	0.0243372	22073	69	<0.001	0.1135579
	Dinucleotide	22244000	345	<0.001	0.01619933	65987	345	<0.001	0.08876235
	Trinucleotide	34086000	1449	<0.001	0.01619454	100910	1071	<0.001	0.1193135
	Tetranucleotide	46124000	5865	<0.001	0.01883876	259800	5865	<0.001	0.1454109
H3K14ac	Nucleotide	7585300	69	<0.001	0.02253996	16731	69	<0.001	0.01197144
	Dinucleotide	17012000	345	<0.001	0.01510283	57183	345	<0.001	0.01010639
	Trinucleotide	26264000	1449	<0.001	0.01516113	103060	1449	<0.001	0.01119785
	Tetranucleotide	35772000	5865	<0.001	0.01770167	182400	5865	<0.001	0.01522907
H3K18ac	Nucleotide	7550800	69	<0.001	0.02269826	4563.3	69	<0.001	0.01032337
	Dinucleotide	16916000	345	<0.001	0.01520117	15231	345	<0.001	0.00860872
	Trinucleotide	26060000	1449	<0.001	0.0152445	27071	1449	<0.001	0.009468009
	Tetranucleotide	35450000	5865	<0.001	0.0177888	45256	5865	<0.001	0.01250661
H3K23ac	Nucleotide	7884200	69	<0.001	0.02273375	7018.9	69	<0.001	0.03026076

	Dinucleotide	17646000	345	<0.001	0.01521959	29733	345	<0.001	0.02902337
	Trinucleotide	27218000	1449	<0.001	0.01527417	83980	1449	<0.001	0.04119723
	Tetranucleotide	37041000	5865	<0.001	0.01782954	190730	5865	<0.001	0.06494022
H3K23me2	Nucleotide	6442300	69	<0.001	0.02278135	79449	69	<0.001	0.01745818
	Dinucleotide	14493000	345	<0.001	0.0152989	204700	345	<0.001	0.01271728
	Trinucleotide	22282000	1449	<0.001	0.01533763	374370	1449	<0.001	0.0141002
	Tetranucleotide	46218000	5865	<0.001	0.01883907	720550	5865	<0.001	0.019865
H3K27ac	Nucleotide	10058000	69	<0.001	0.02440445	1327.7	54	<0.001	0.01985795
	Dinucleotide	22276000	345	<0.001	0.01624317	3342.8	270	<0.001	0.01419714
	Trinucleotide	34125000	1449	<0.001	0.01623601	7290.1	1134	<0.001	0.01928499
	Tetranucleotide	46160000	5865	<0.001	0.0188838	18297	4590	<0.001	0.03078512
H3K27me3	Nucleotide	10052000	69	<0.001	0.02435711	1750.1	69	<0.001	0.06224492
	Dinucleotide	22264000	345	<0.001	0.01621153	8013.2	345	<0.001	0.06022942
	Trinucleotide	34112000	1449	<0.001	0.01620582	22493	1449	<0.001	0.08241953
	Tetranucleotide	46153000	5865	<0.001	0.01885075	50705	5865	<0.001	0.1251495
H3K36me3	Nucleotide	10054000	69	<0.001	0.02436328	1249	69	<0.001	0.05107377
	Dinucleotide	22267000	345	<0.001	0.01621573	7368.2	345	<0.001	0.05637531
	Trinucleotide	34117000	1449	<0.001	0.01620988	22085	1449	<0.001	0.08013865
	Tetranucleotide	46159000	5865	<0.001	0.01885508	50380	5865	<0.001	0.1230405
H3K56ac	Nucleotide	7297300	69	<0.001	0.02276179	62668	69	<0.001	0.01947377
	Dinucleotide	16346000	345	<0.001	0.01524534	204440	345	<0.001	0.01583134
	Trinucleotide	25151000	1449	<0.001	0.01528214	489610	1449	<0.001	0.01991447
	Tetranucleotide	34126000	5865	<0.001	0.0178132	1127500	5865	<0.001	0.03041842
H3K79me1	Nucleotide	8099300	69	<0.001	0.02289817	10838	69	<0.001	0.020081
	Dinucleotide	18077000	345	<0.001	0.01530745	42645	345	<0.001	0.0183584
	Trinucleotide	27837000	1449	<0.001	0.01534884	122370	1449	<0.001	0.02593187
	Tetranucleotide	37831000	5865	<0.001	0.01790355	312210	5865	<0.001	0.04278113
H3K79me2	Nucleotide	7937500	69	<0.001	0.02282001	50729	69	<0.001	0.01395807
	Dinucleotide	17756000	345	<0.001	0.01527137	135350	345	<0.001	0.01028522
	Trinucleotide	27373000	1449	<0.001	0.01532011	248100	1449	<0.001	0.01134579
	Tetranucleotide	37227000	5865	<0.001	0.01787507	472360	5865	<0.001	0.01579449
H3T11ph	Nucleotide	1236100	69	<0.001	0.02116811	1236100	69	<0.001	0.02116811
	Dinucleotide	2770600	345	<0.001	0.01434907	2770600	345	<0.001	0.01434907
	Trinucleotide	4221300	1449	<0.001	0.0144861	4221300	1449	<0.001	0.0144861
	Tetranucleotide	5723700	5865	<0.001	0.01708886	5723700	5865	<0.001	0.01708886
H4K5ac	Nucleotide	7839600	69	<0.001	0.02346504	7839600	69	<0.001	0.02346504
	Dinucleotide	17513000	345	<0.001	0.01569392	17513000	345	<0.001	0.01569392

	Trinucleotide	26869000	1449	<0.001	0.01570743	26869000	1449	<0.001	0.01570743
	Tetranucleotide	36369000	5865	<0.001	0.0182854	880560	5865	<0.001	0.01597168
H4K8ac	Nucleotide	6773900	69	<0.001	0.0220494	2167.1	69	<0.001	0.01534723
	Dinucleotide	15240000	345	<0.001	0.01479986	8737.9	345	<0.001	0.01409647
	Trinucleotide	23529000	1449	<0.001	0.01486008	19375	1449	<0.001	0.01735714
	Tetranucleotide	32062000	5865	<0.001	0.01735739	48831	5865	<0.001	0.02822063
H4K12ac	Nucleotide	2366200	69	<0.001	0.01764977	500900	69	<0.001	0.01352
	Dinucleotide	5546300	345	<0.001	0.01210017	1273800	345	<0.001	0.009666201
	Trinucleotide	8719000	1449	<0.001	0.01226792	2082100	1449	<0.001	0.01000513
	Tetranucleotide	12066000	5865	<0.001	0.01445066	3010800	5865	<0.001	0.01206156
H4K20me1	Nucleotide	6190100	69	<0.001	0.02216194	42804	69	<0.001	0.0155282
	Dinucleotide	13970000	345	<0.001	0.01490241	125030	345	<0.001	0.01193584
	Trinucleotide	21603000	1449	<0.001	0.01497875	257350	1449	<0.001	0.01390832
	Tetranucleotide	29456000	5865	<0.001	0.01750593	546420	5865	<0.001	0.02038286
H4K91ac	Nucleotide	6653000	69	<0.001	0.02194719	10784	69	<0.001	0.01419887
	Dinucleotide	14995000	345	<0.001	0.01474491	36701	345	<0.001	0.01196824
	Trinucleotide	23185000	1449	<0.001	0.01481648	74993	1449	<0.001	0.01412876
	Tetranucleotide	31613000	5865	<0.001	0.01731289	165280	5865	<0.001	0.02145625

Chi-square test value for histone union and histone intersection. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p value; Cramer's V matches measure of association Cramer's V.

Annex 7 – Comparing to the each histone with control

		Union			
		X^2	df	p-value	Cramer's V
H2AZ	Nucleotide	8969.3	3	<0.001	0.001256411
	Dinucleotide	29912	15	<0.001	0.002294481
	Trinucleotide	54176	63	<0.001	0.003087987
	Tetranucleotide	91899	255	<0.001	0.004021933
H2AK5ac	Nucleotide	464.68	3	<0.001	0.0002982823
	Dinucleotide	11310	15	<0.001	0.001472304
	Trinucleotide	26998	63	<0.001	0.002275844
	Tetranucleotide	58741	255	<0.001	0.003358645
H2AK9ac	Nucleotide	252440	3	<0.001	0.01409013
	Dinucleotide	596720	15	<0.001	0.0218812
	Trinucleotide	904450	63	<0.001	0.02721546
	Tetranucleotide	1211300	255	<0.001	0.03182611
H2BK5ac	Nucleotide	663.61	3	<0.001	0.0003576462
	Dinucleotide	11897	15	<0.001	0.001515236
	Trinucleotide	28053	63	<0.001	0.002328113
	Tetranucleotide	60067	255	<0.001	0.003408693
H2BK12ac	Nucleotide	1044.7	3	<0.001	0.0004615974
	Dinucleotide	12950	15	<0.001	0.001626103
	Trinucleotide	29959	63	<0.001	0.002474678
	Tetranucleotide	64201	255	<0.001	0.003624668
H2BK15ac	Nucleotide	1287	3	<0.001	0.0004931837
	Dinucleotide	13187	15	<0.001	0.001579155
	Trinucleotide	29794	63	<0.001	0.002374258
	Tetranucleotide	62123	255	<0.001	0.003429333
H2BK20ac	Nucleotide	461.94	3	<0.001	0.0003133779
	Dinucleotide	12502	15	<0.001	0.001631332
	Trinucleotide	29478	63	<0.001	0.00250655

	Tetranucleotide	62291	255	<0.001	0.003646035
H2BK120ac	Nucleotide	32.93	3	<0.001	8.071568e-05
	Dinucleotide	10381	15	<0.001	0.001433721
	Trinucleotide	26085	63	<0.001	0.002273574
	Tetranucleotide	58877	255	<0.001	0.003417123
H3K4ac	Nucleotide	5.044	3	<0.001	3.146293e-05
	Dinucleotide	10499	15	<0.001	0.001435971
	Trinucleotide	26114	63	<0.001	0.002265604
	Tetranucleotide	58374	255	<0.001	0.003388623
H3K4me1	Nucleotide	8874.5	3	<0.001	0.001252109
	Dinucleotide	29748	15	<0.001	0.002292513
	Trinucleotide	54002	63	<0.001	0.003088855
	Tetranucleotide	91853	255	<0.001	0.004028558
H3K4me2	Nucleotide	3536.2	3	<0.001	0.0008119262
	Dinucleotide	18444	15	<0.001	0.001854808
	Trinucleotide	37575	63	<0.001	0.002648153
	Tetranucleotide	71300	255	<0.001	0.0036489
H3K4me3	Nucleotide	8933	3	<0.001	0.001256336
	Dinucleotide	29887	15	<0.001	0.002298042
	Trinucleotide	54205	63	<0.001	0.00309493
	Tetranucleotide	92103	255	<0.001	0.004034413
H3K9ac	Nucleotide	8945.7	3	<0.001	0.001260505
	Dinucleotide	30059	15	<0.001	0.002310687
	Trinucleotide	54655	63	<0.001	0.003115917
	Tetranucleotide	93103	255	<0.001	0.004066966
H3K9me1	Nucleotide	22465	3	<0.001	0.002822762
	Dinucleotide	66619	15	<0.001	0.004867332
	Trinucleotide	114030	63	<0.001	0.00637644
	Tetranucleotide	179300	255	<0.001	0.008006405

H3K9me3	Nucleotide	8944.6	3	<0.001	0.001255921
	Dinucleotide	29882	15	<0.001	0.002295626
	Trinucleotide	54144	63	<0.001	0.003090135
	Tetranucleotide	91891	255	<0.001	0.004025783
H3K14ac	Nucleotide	273.29	3	<0.001	0.0002338814
	Dinucleotide	11340	15	<0.001	0.001507276
	Trinucleotide	27424	63	<0.001	0.002344976
	Tetranucleotide	60240	255	<0.001	0.003477008
H3K18ac	Nucleotide	1213.1	3	<0.001	0.0004973358
	Dinucleotide	12927	15	<0.001	0.001624284
	Trinucleotide	30440	63	<0.001	0.002493756
	Tetranucleotide	66355	255	<0.001	0.003683706
H3K23ac	Nucleotide	23.462	3	<0.001	6.779707e-05
	Dinucleotide	10526	15	<0.001	0.001436914
	Trinucleotide	26032	63	<0.001	0.002261107
	Tetranucleotide	57795	255	<0.001	0.003371208
H3K23me2	Nucleotide	4732.6	3	<0.001	0.001066978
	Dinucleotide	21307	15	<0.001	0.002266581
	Trinucleotide	41982	63	<0.001	0.003185364
	Tetranucleotide	91899	255	<0.001	0.004021933
H3K27ac	Nucleotide	8831.8	3	<0.001	0.00125044
	Dinucleotide	29686	15	<0.001	0.002292614
	Trinucleotide	53938	63	<0.001	0.003090396
	Tetranucleotide	91845	255	<0.001	0.004032803
H3K27me3	Nucleotide	8922.9	3	<0.001	0.001254797
	Dinucleotide	29849	15	<0.001	0.002295087
	Trinucleotide	54124	63	<0.001	0.003090567
	Tetranucleotide	91931	255	<0.001	0.004027942
H3K36me3	Nucleotide	8905.2	3	<0.001	0.001253761
	Dinucleotide	29800	15	<0.001	0.002293578

	Trinucleotide	54049	63	<0.001	0.003088949
	Tetranucleotide	91851	255	<0.001	0.004026886
H3K56ac	Nucleotide	285.29	3	<0.001	0.0002460015
	Dinucleotide	11985	15	<0.001	0.001595525
	Trinucleotide	28666	63	<0.001	0.00246921
	Tetranucleotide	60614	255	<0.001	0.003592993
H3K79me1	Nucleotide	465.71	3	<0.001	0.0003001791
	Dinucleotide	11360	15	<0.001	0.001483417
	Trinucleotide	27329	63	<0.001	0.002302114
	Tetranucleotide	59456	255	<0.001	0.003397546
H3K79me2	Nucleotide	93.798	3	<0.001	0.0001356138
	Dinucleotide	10702	15	<0.001	0.001449282
	Trinucleotide	26253	63	<0.001	0.002271053
	Tetranucleotide	57912	255	<0.001	0.003374742
H3T11ph	Nucleotide	22900	3	<0.001	0.004938626
	Dinucleotide	63722	15	<0.001	0.00833877
	Trinucleotide	106090	63	<0.001	0.01089437
	Tetranucleotide	161690	255	<0.001	0.01362157
H4K5ac	Nucleotide	281.9	3	<0.001	0.0002432175
	Dinucleotide	11665	15	<0.001	0.001565502
	Trinucleotide	28709	63	<0.001	0.002457397
	Tetranucleotide	62241	255	<0.001	0.00362042
H4K8ac	Nucleotide	4612.3	3	<0.001	0.000994479
	Dinucleotide	20489	15	<0.001	0.002097331
	Trinucleotide	41790	63	<0.001	0.002997183
	Tetranucleotide	81665	255	<0.001	0.004192455
H4K12ac	Nucleotide	219740	3	<0.001	0.009280546
	Dinucleotide	522470	15	<0.001	0.01432888
	Trinucleotide	794490	63	<0.001	0.01769254
	Tetranucleotide	1076700	255	<0.001	0.02062374

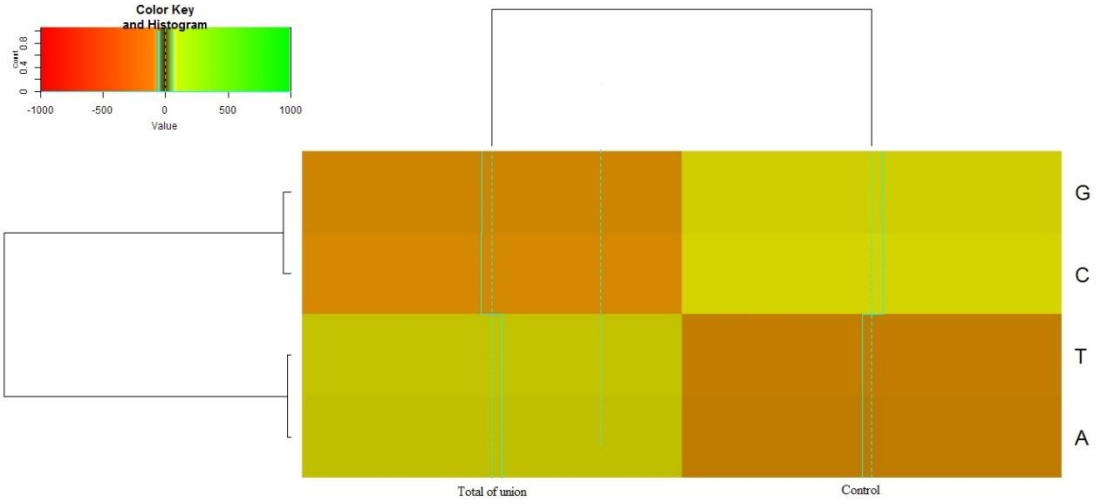
H4K20me1	Nucleotide	6182.6	3	<0.001	0.001210335
	Dinucleotide	25005	15	<0.001	0.002436217
	Trinucleotide	48282	63	<0.001	0.003388234
	Tetranucleotide	87970	255	<0.001	0.004577497
H4K91ac	Nucleotide	4812.4	3	<0.001	0.001020235
	Dinucleotide	21666	15	<0.001	0.00216619
	Trinucleotide	43102	63	<0.001	0.003057366
	Tetranucleotide	81702	255	<0.001	0.004212164

Chi-square test value for comparing to the each histone with the control. X^2 matches chi-square test; df matches degrees of freedom; p-value matches p value; Cramer's V matches measure of association Cramer's V.

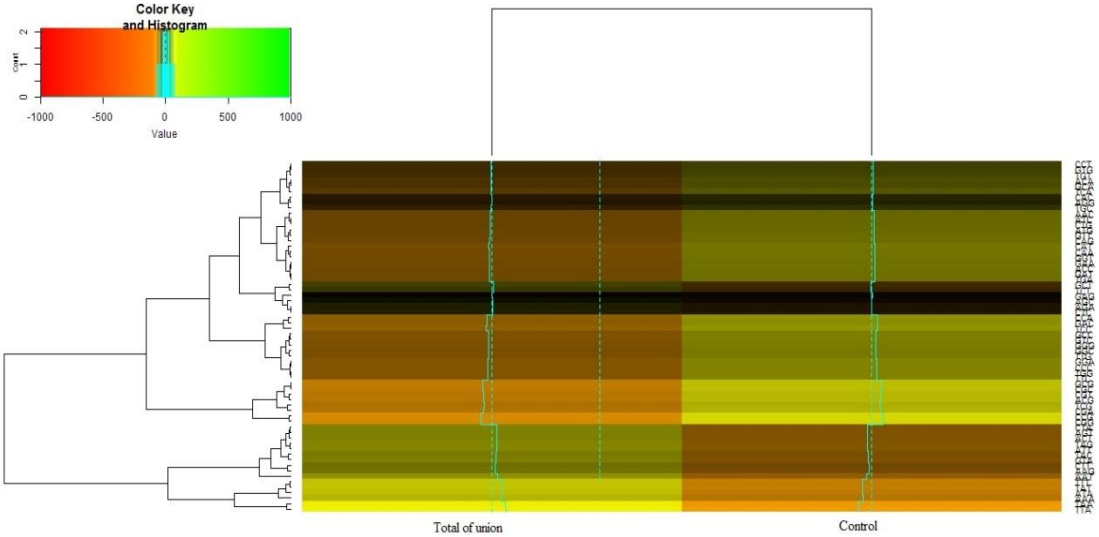
Annex 8 – Heatmaps

Heatmaps corresponding to the analysis the comparing the total of genomes with to control, to understand if the context of epigenetic marking the chromosomes is equal to the epigenetic marking context in control.

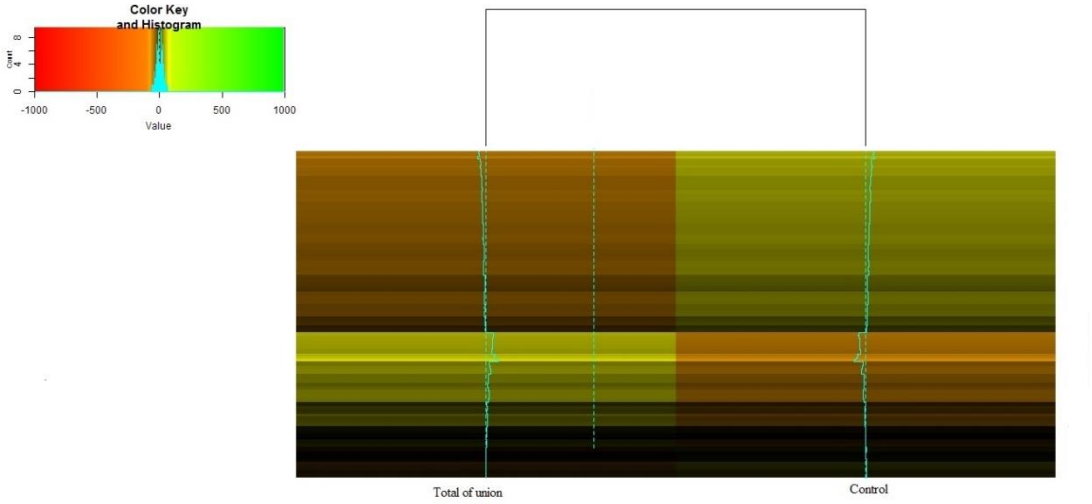
- Analysis union-control for nucleotide:



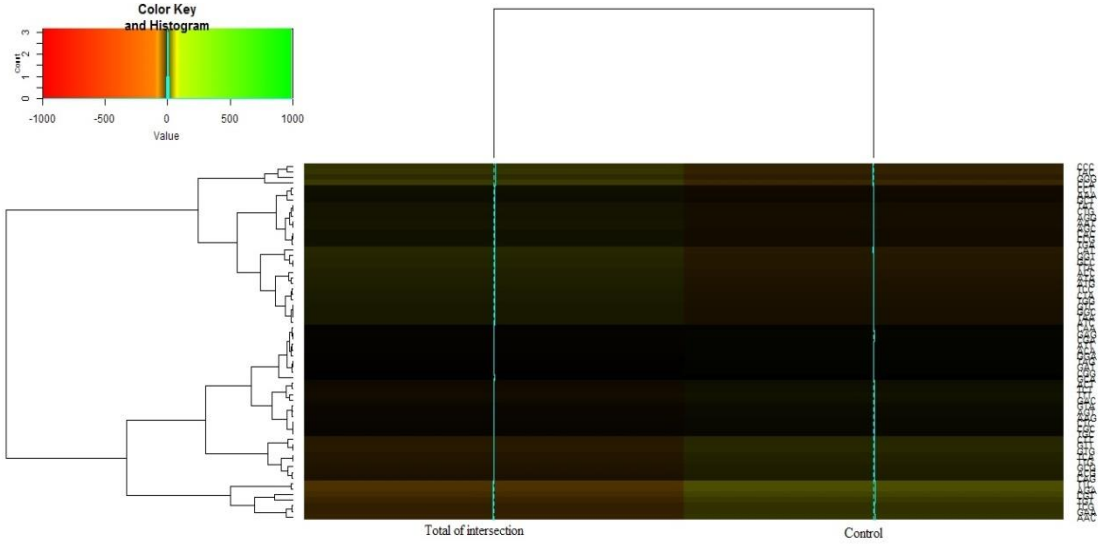
- Analysis union-control for trinucleotide:



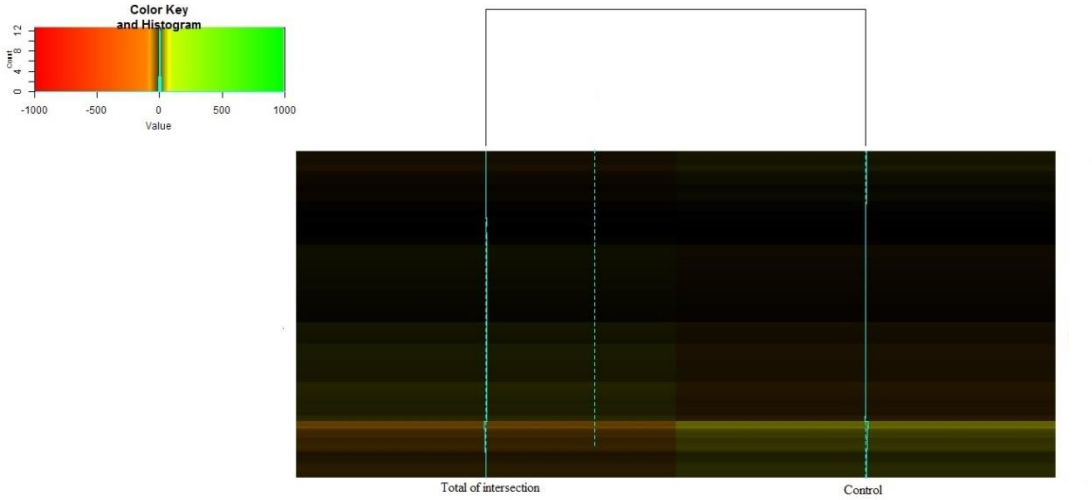
- Analysis union-control for tetranucleotide:



- Analysis inter-control for trinucleotide:

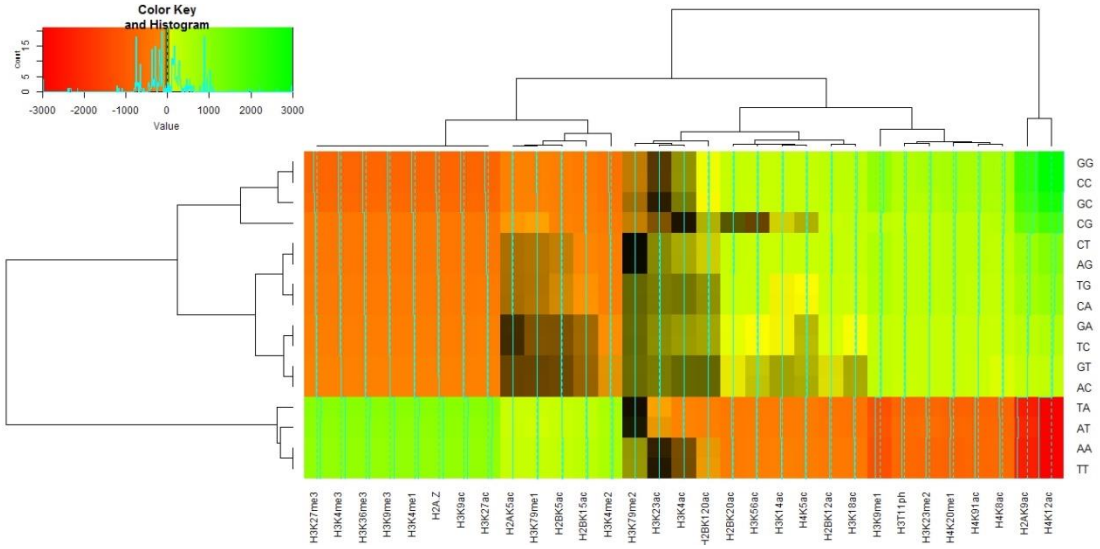


- Analysis inter-control for tetranucleotide:

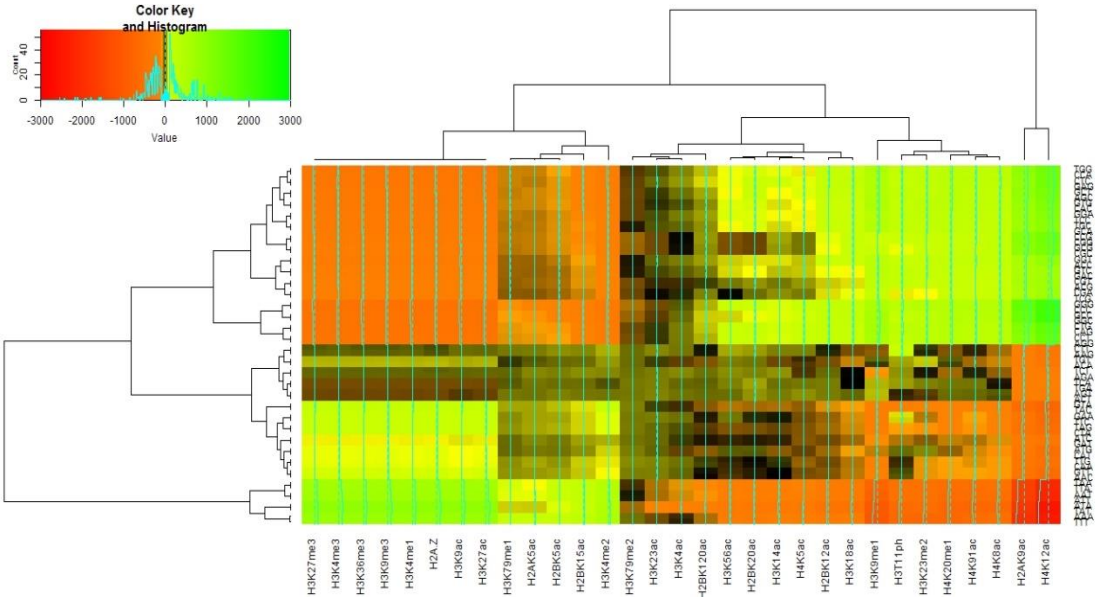


Heatmaps that we see if the context of epigenetic marking is homogeneous among histones, for histone union and histone intersection.

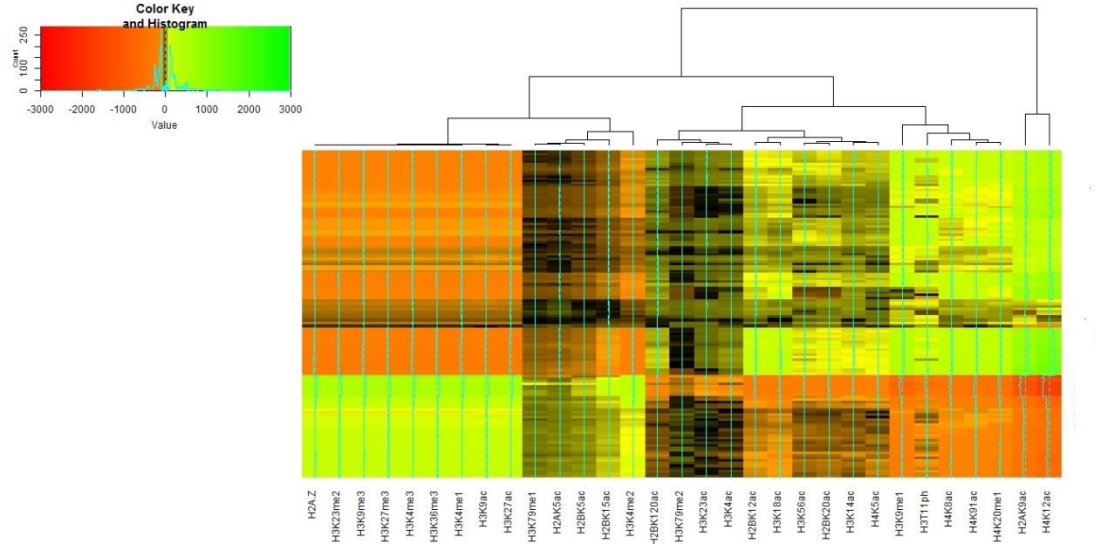
- Histone union for dinucleotide:



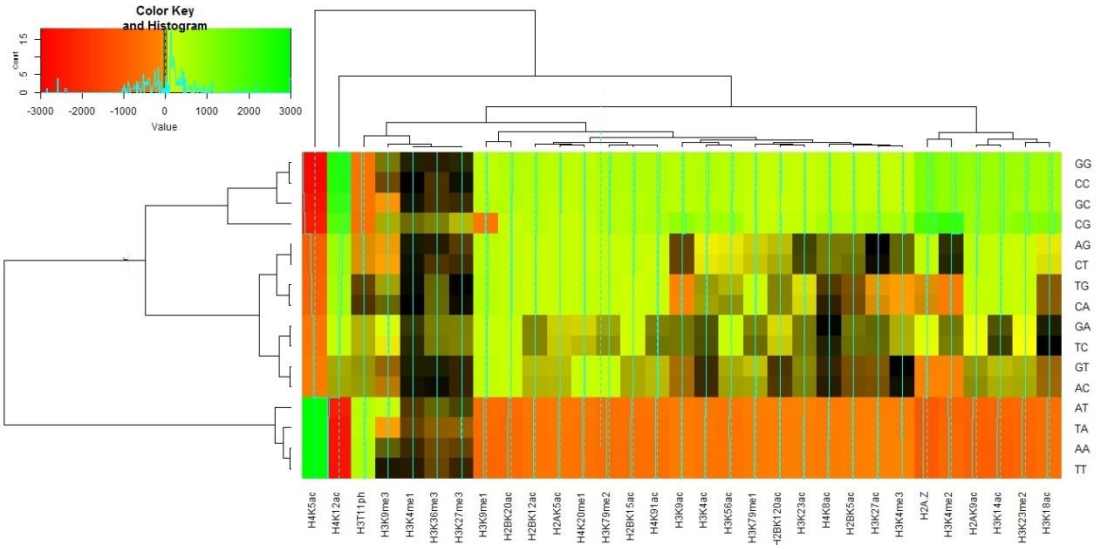
- Histone union for trinucleotide:



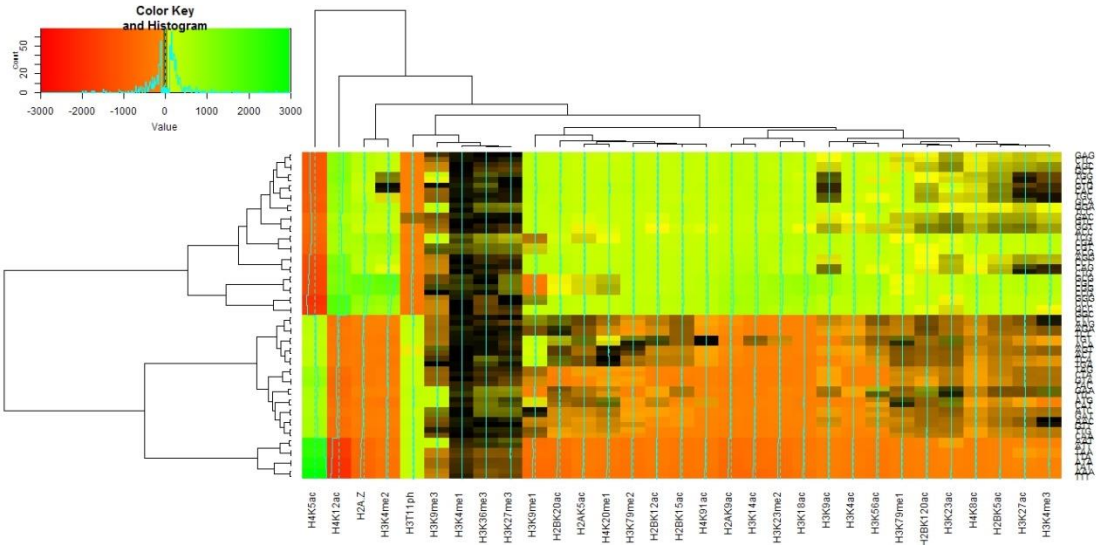
- Histone union for tetranucleotide:



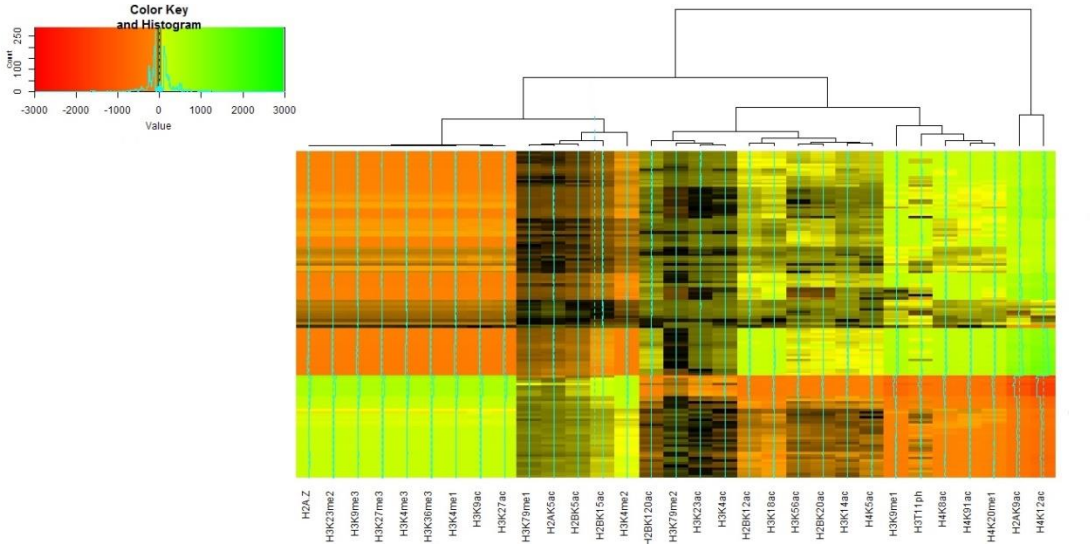
- Histone intersection for dinucleotide:



- Histone intersection for trinucleotide:

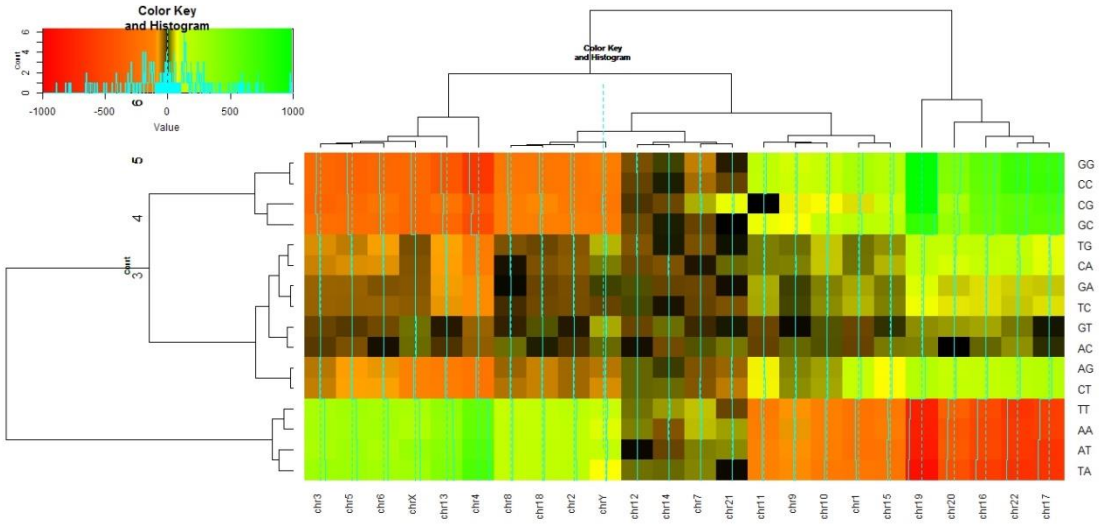


- Histone intersection for tetranucleotide:

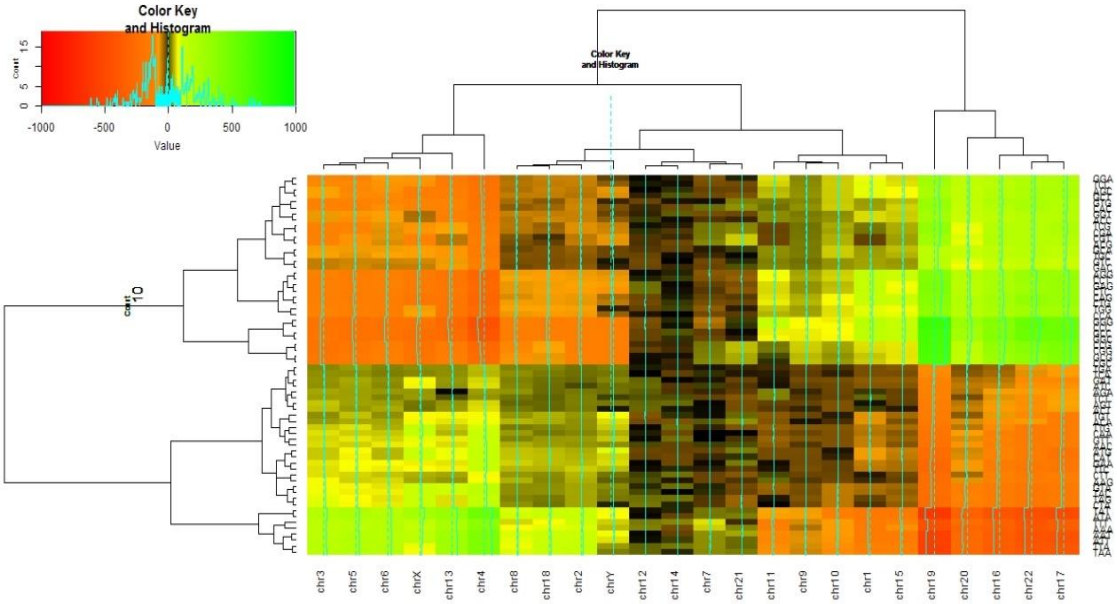


Heatmaps corresponding to the analysis that evaluates the occurrence epigenetic marking in chromosomes is homogeneous, in the marking zone and the non marking zone.

- Global union for dinucleotide:



- Global union for trinucleotide:



- Global union for tetranucleotide: